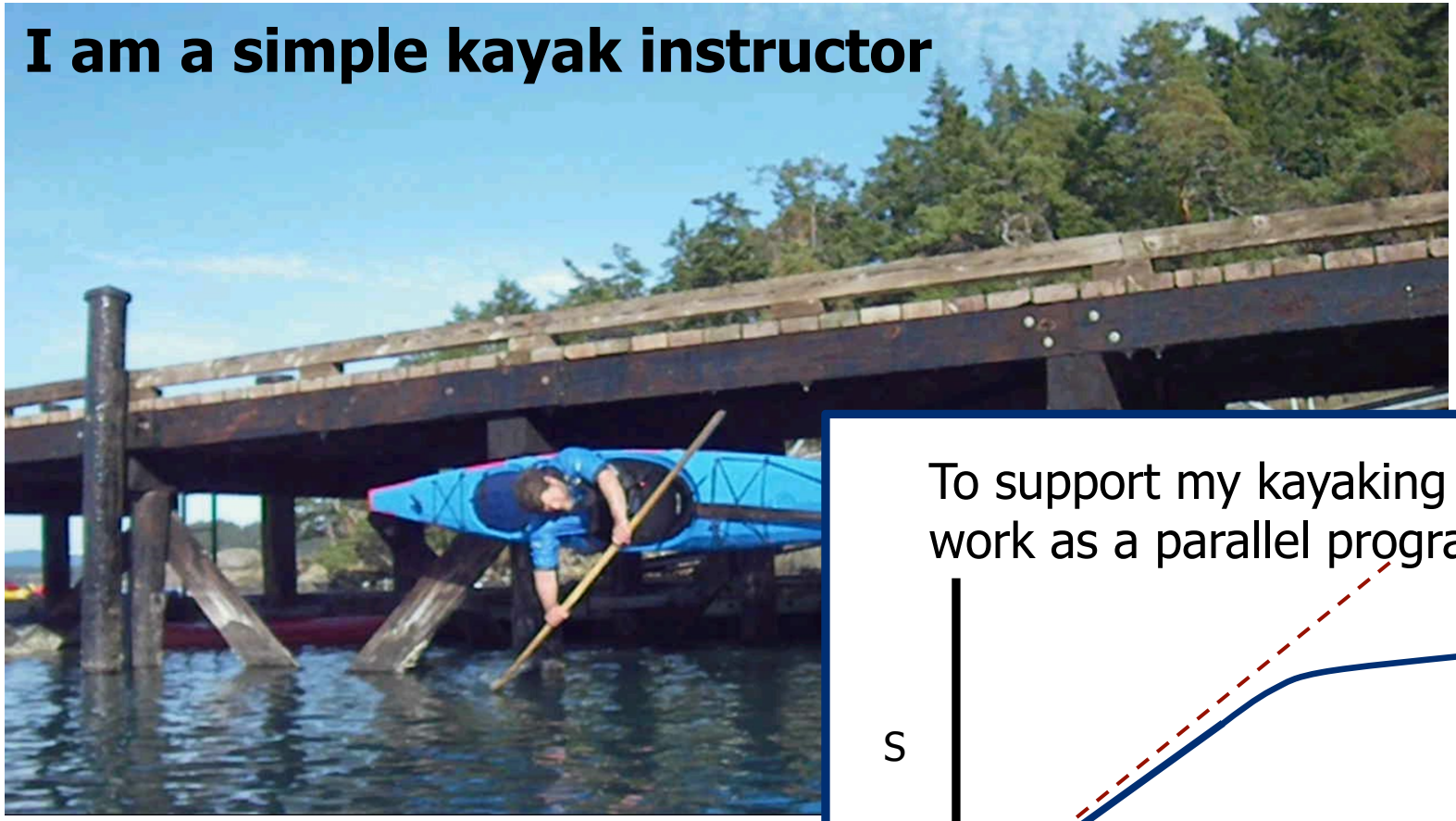
Big Data: Lessons learned by a confused novice

Tim Mattson,
Intel Parallel Computing Lab

Introduction



To support my kayaking habit I work as a parallel programmer

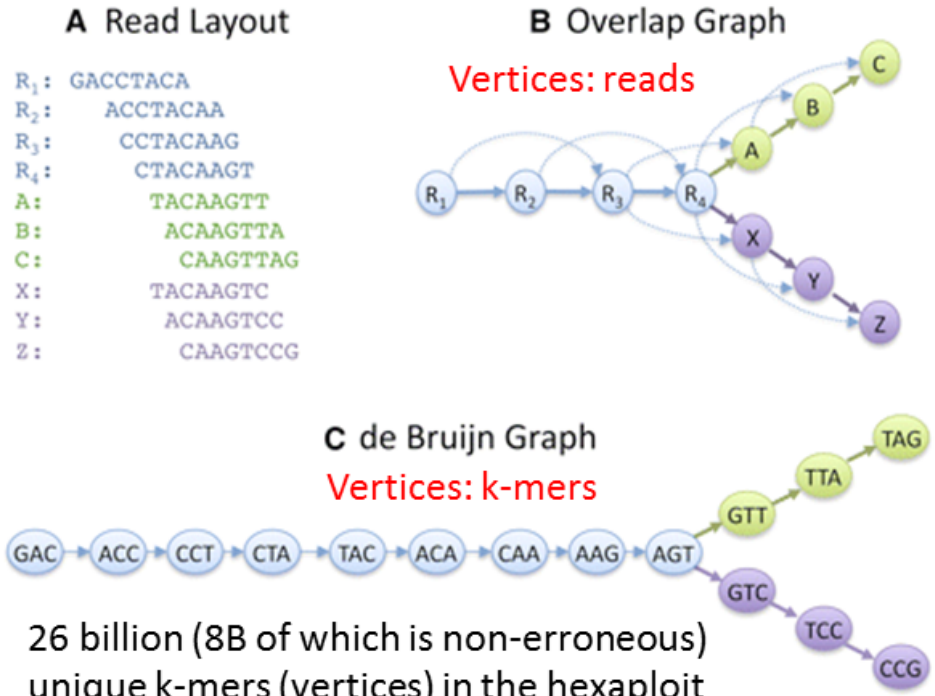
Which means I know how to turn math into lines on a speedup plot

My Journey into Big Data

I spent a few years working on Parallel Graph algorithms

matter in Applied Math

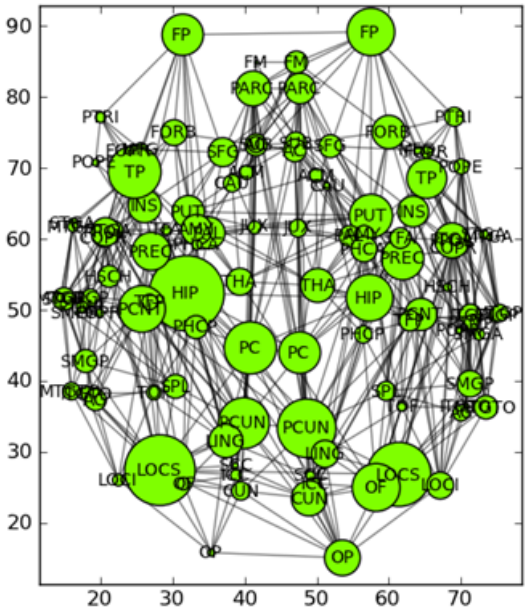
genome assembly



26 billion (8B of which is non-erroneous) unique k-mers (vertices) in the hexaploit wheat genome W7984 for k=51

Schatz et al. (2010) Perspective: Assembly of Large Genomes w/2nd-Gen Seq. Genome Res. (figure reference)

Graph Theoretical analysis of Brain Connectivity

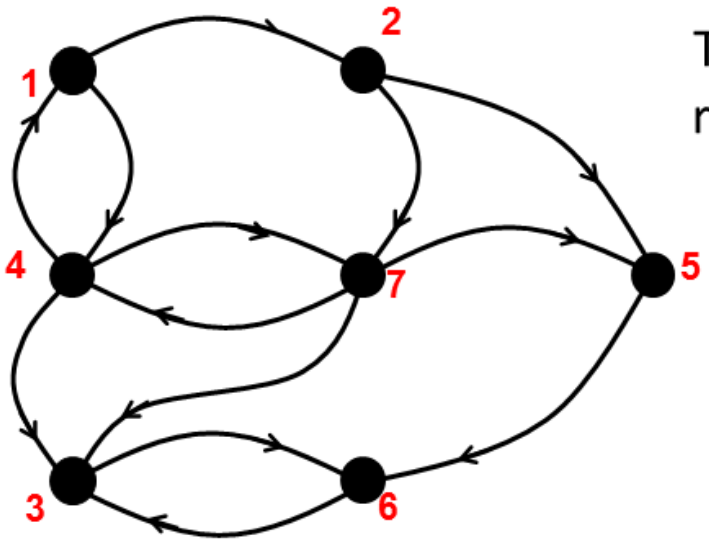


Potentially millions of neurons and billions of edges with developing technologies

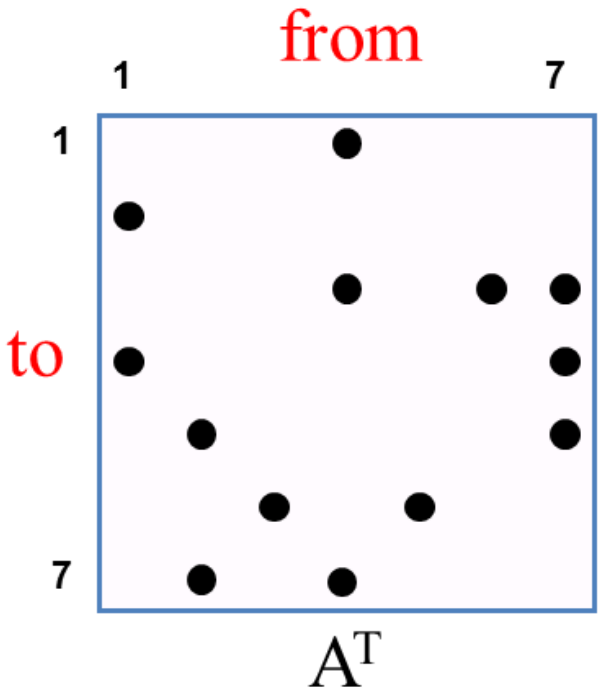
My Journey into Big Data

I spent a few years working on Parallel Graph algorithms

Graphs in the Language of Linear Algebra



These two diagrams are equivalent representations of the same graph.

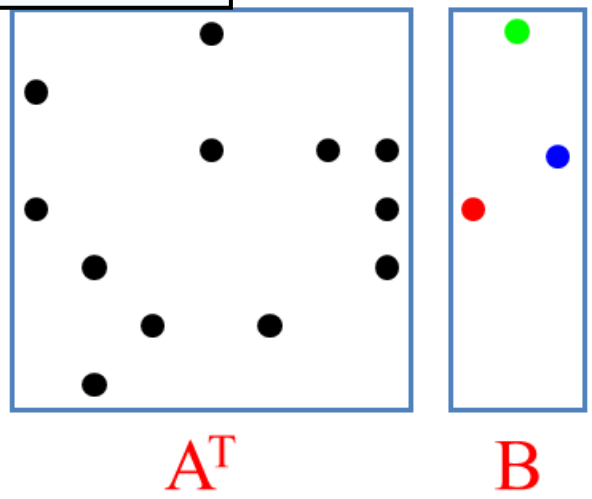


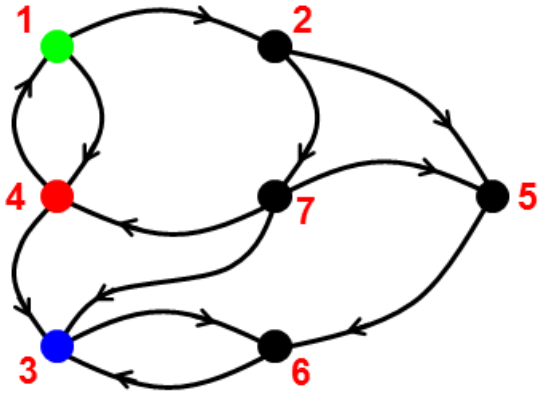
A = the adjacency matrix …
Elements nonzero when vertices are adjacent

My Journey into Big Data

I spent a few years working on Parallel Graph algorithms

Multiple-source breadth-first search





- Sparse array representation => space efficient
- Sparse matrix-matrix multiplication => work efficient
- Three possible levels of parallelism: searches, vertices, edges
- Highly-parallel implementation for Betweenness Centrality*

*: A measure of influence in graphs, based on shortest paths

Source: Aydin Buluc, et. al. The Graph BLAS effort and its implications for Exascale

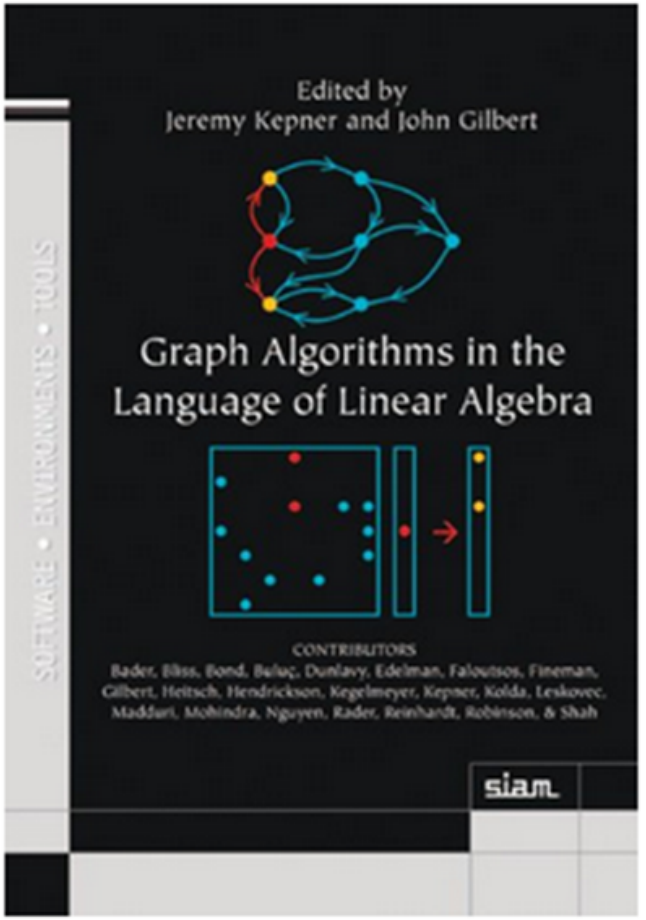
My Journey into Big Data

I spent a few years working on Parallel Graph algorithms

A "new" conceptual framework for graph algorithms

- Jeremy Kepner, John Gilbert and friends showed how you can build the full range of parallel graph algorithms on top of a linear algebra foundation.

- Aydin Buluc working with John Gilbert at UCSB showed how to make them fast with the combinatorial BLAS.



My Journey into Big Data

I spent a few years working on Parallel Graph algorithms

The Graph BLAS effort

Standards for Graph Algorithm Primitives

Tim Mattson (Intel Corporation), David Bader (Georgia Institute of Technology), Jon Berry (Sandia National Laboratory), Aydin Buluc (Lawrence Berkeley National Laboratory), Jack Dongarra (University of Tennessee), Christos Faloutsos (Carnegie Melon University), John Feo (Pacific Northwest National Laboratory), John Gilbert (University of California at Santa Barbara), Joseph Gonzalez (University of California at Berkeley), Bruce Hendrickson (Sandia National Laboratory), Jeremy Kepner (Massachusetts Institute of Technology), Charles Leiserson (Massachusetts Institute of Technology), Andrew Lumsdaine (Indiana University), David Padua (University of Illinois at Urbana-Champaign), Stephen Poole (Oak Ridge National Laboratory), Steve Reinhardt (Cray Corporation), Mike Stonebraker (Massachusetts Institute of Technology), Steve Wallach (Convey Corporation), Andrew Yoo (Lawrence Livermore National Laboratory)

Abstract-- It is our view that the state of the art in constructing a large collection of graph algorithms in terms of linear algebraic operations is mature enough to support the emergence of a standard set of primitive building blocks. This paper is a position paper defining the problem and announcing our intention to launch an open effort to define this standard.

- The Graph BLAS Forum: http://istc-bigdata.org/GraphBlas/

My Journey into Big Data

I spent a few years working on Parallel Graph algorithms

The Graph BLAS effort

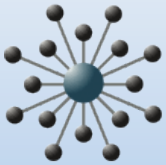
Standards for Graph Algorithm Primitives

Tim Mattson (Intel Corporation), David Bader (Georgia Institute of Technology), Jon Berry (Sandia National Laboratory), Aydin Buluc (Lawrence Berkeley National Laboratory), Jack Dongarra (University of Tennessee), Christos Faloutsos (Carnegie Melon University), John Feo (Pacific Northwest National Laboratory), John Gilbert (University of California at Santa Barbara), Joseph Gonzalez (University of California at Berkeley), Bruce Hendrickson (Sandia National Laboratory), Jeremy Kepner (Massachusetts Institute of Technology), Charles Leiserson (Massachusetts Institute of Technology), Andrew Lumsdaine (Indiana University), David Padua (University of Illinois at Urbana-Champaign), Stephen Poole (Oak Ridge National Laboratory), Steve Reinhardt (Cray Corporation), Mike Stonebraker (Massachusetts Institute of Technology), Steve Wallach (Convey Corporation), Andrew Yoo (Lawrence Livermore National Laboratory)

Abstract-- It is our view that the state of the art in constructing a large collection of graph algorithms in terms of linear algebr
support the emergence of a standard set
a position paper defining the problem and
open effort to define this standard.

- The Graph BLAS Forum: http://istc-bigd

… and since graphs are heavily used in Big Data, that means I'm a "Big Data Expert"



ISTC BIG DATA

… 2 years ago Intel made me responsible for a big data research center

ISTC Research themes

- Data Analytics & Processing Platforms
- Scalable Math and Algorithms
- Visualization
- Architecture

I was WAY out of my league … but I'm slowly coming up to speed. I've learned 4 key lessons over the last year!

intel MIT BROWN Carnegie Mellon University W UT Portland State UNIVERSITY

Lesson 1: Hype Abounds

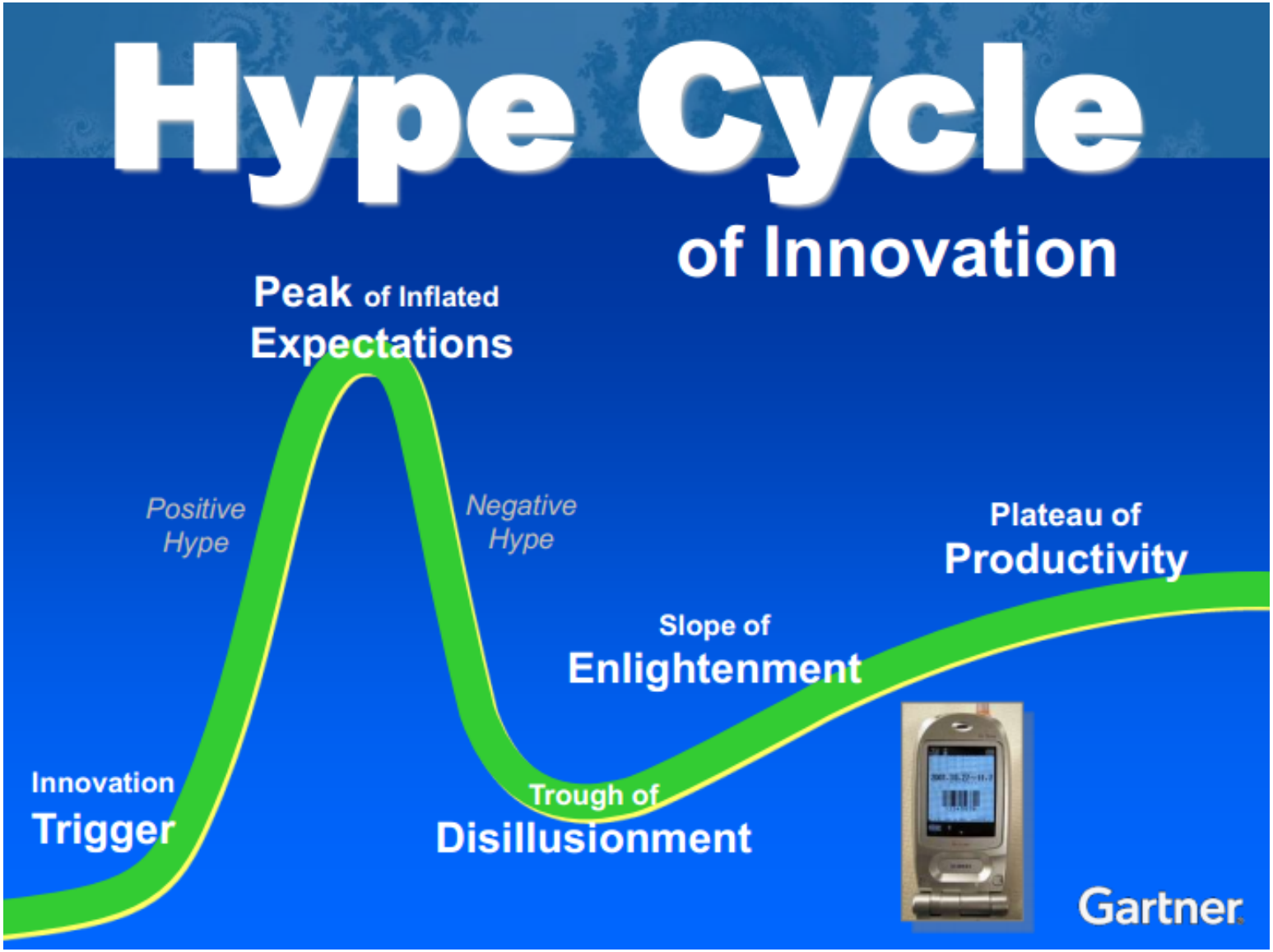
- This truly is the most overly hyped field I've ever touched.

BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK … big-data analytics is revolutionizing the way we see and process the world … compare its consequences to those of the Gutenberg printing press.

The Big Data Revolution will be bigger than the internet.

Big Data has ended privacy as we know it … Big Data will let Big-Government/Big-Business control your life.

Understanding the structure of hype



Source: http://www.calisto.bg/userfiles/file/Mastering%20Hype%20Cycle%20-%20Sofia.pdf (collected Oct 12, 2014)

Understanding the structure of hype

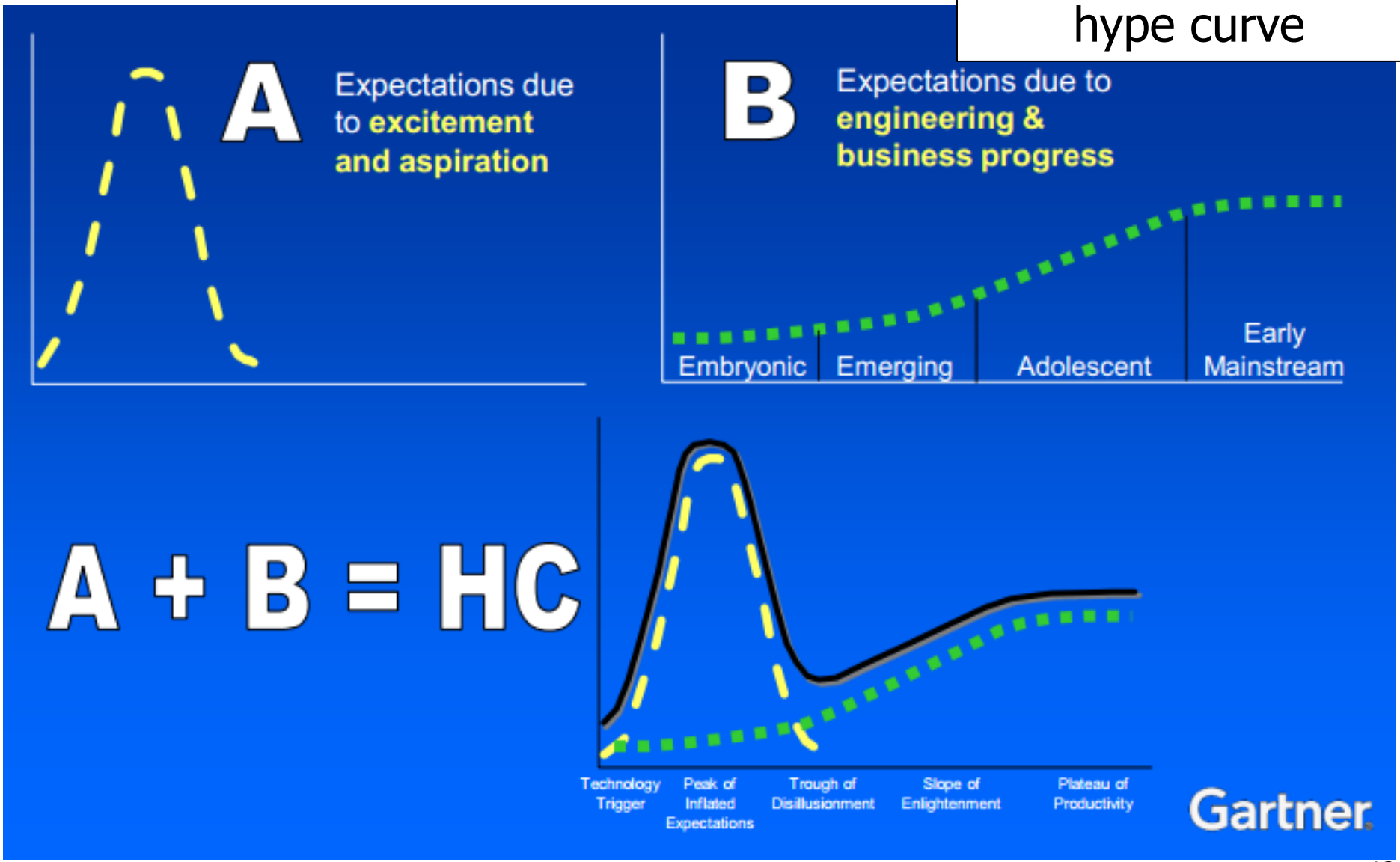
This pattern is very common

Gartner

Source: http://www.calisto.bg/userfiles/file/Mastering%20Hype%20Cycle%20-%20Sofia.pdf (collected Oct 12, 2014)

Understanding the structure of hype

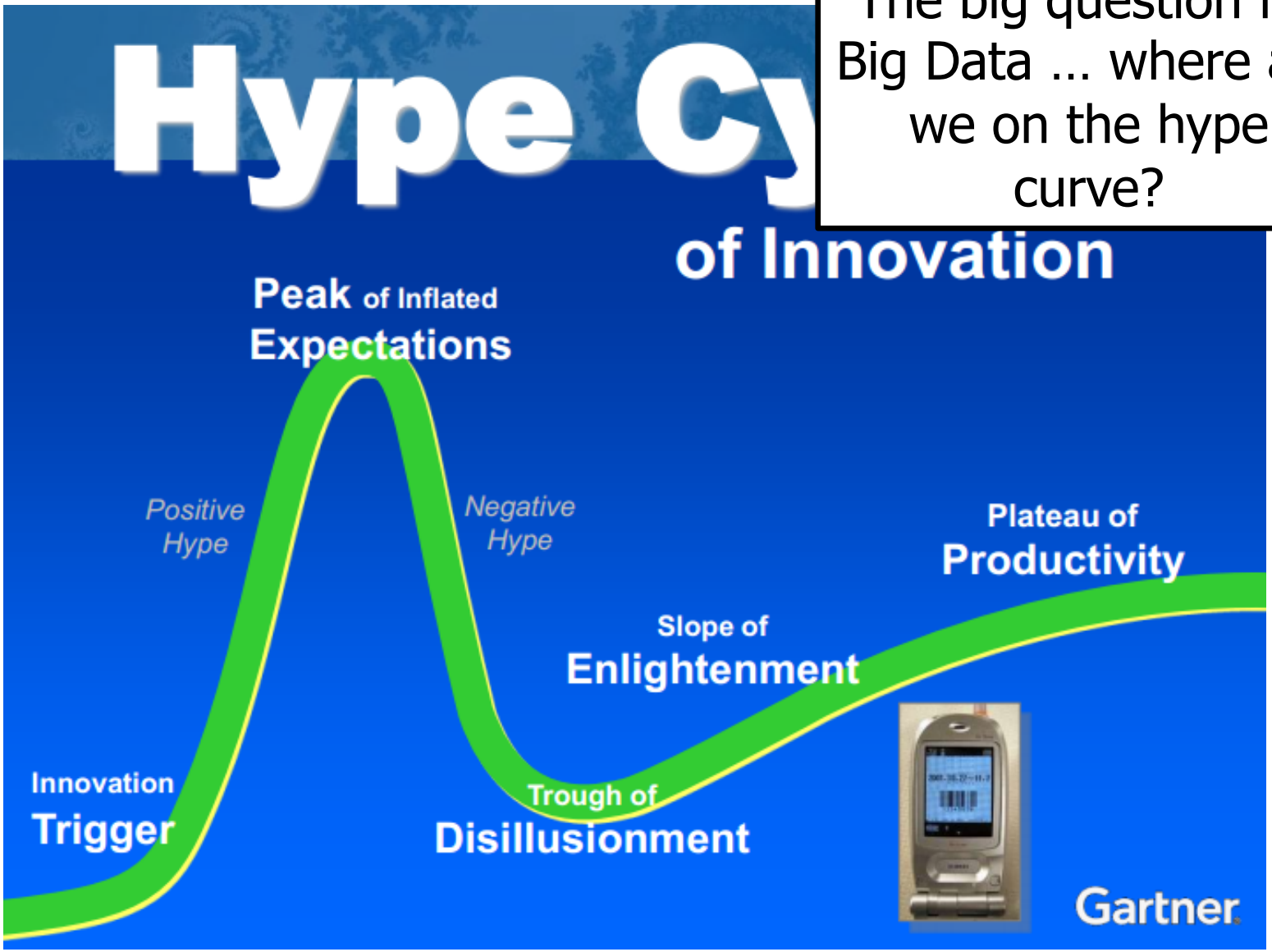
The origin of the hype curve



Source: http://www.calisto.bg/userfiles/file/Mastering%20Hype%20Cycle%20-%20Sofia.pdf (collected Oct 12, 2014)

Understanding the structure of hype

The big question for Big Data … where are we on the hype curve?



Source: http://www.calisto.bg/userfiles/file/Mastering%20Hype%20Cycle%20-%20Sofia.pdf (collected Oct 12, 2014)

Lesson 2: Big Data isn't that big (so far)

- Most Big Data data-sets are not that Big compared to modern storage technologies.
- Consider the following
  - 300 Million people in the U.S..
  - A genome has 3 billion base pairs, but we only need to store the differences relative to a reference genome … maybe 1 percent.
  - The genetic code uses four nucleotides … i.e. a 2 bit code.
  - So I can store key genomic data on entire US population in:
    - 3*10^8 people * 3*10^9 base-pairs * 0.01*2 bit * 1 byte/8 bit ≈ **2 Petabytes**
  - And there is so much similarity between genomes, this data could be compressed an additional one or two orders of magnitude.

So if it's not the size, what is big data really all about?

Big Data: It's more than the size

The Need For New Data Platforms

- Big Data ➔ "Volume, Velocity, Variety"
- Velocity, Variety Present the Biggest Challenges
- Examples:
  - **Velocity:** Transaction processing: how to handle hundreds of millions of transactions per day?
    - How to exploit growth of main memory, manycore?
  - **Variety:** Array Databases: increasingly people want machine learning & predictive analytics on data
    - These algorithms are expressed on arrays
    - How to take advantage of manycore platforms?
  - Graphs, networks, etc…

Big Data: It's more than the size

The Need For New Data Platforms

- Big Data ➔ "Volume, Velocity, Variety"
- Velocity, Variety Present the Biggest Challenges

Response … automation; i.e. You can't do Big Data "by hand":

- Automatically Stream data into your data store system
- Automatically condition data so you can use it
- Automatically generate classifiers and find trends
- Automatically discover "actionable knowledge"

– Graphs, networks, etc…

Lesson 3: Move queries, not data

- A motivating example:
  - Use fMRI data to connect the mind to the brain … requires processing at human-interactive speeds (~ 1 Sec.).
  - Assume a typical data center:
    - High speed network (2 GB/s).
    - Separate data and computer servers.
  - Data set is 480 GB. Do the math:
    - 480 GB * (1 sec/2 GB) →4 Min
  - But it's even worse …
    - We actual want to do correlations across hundreds of subjects … 400 minutes just to move data around!

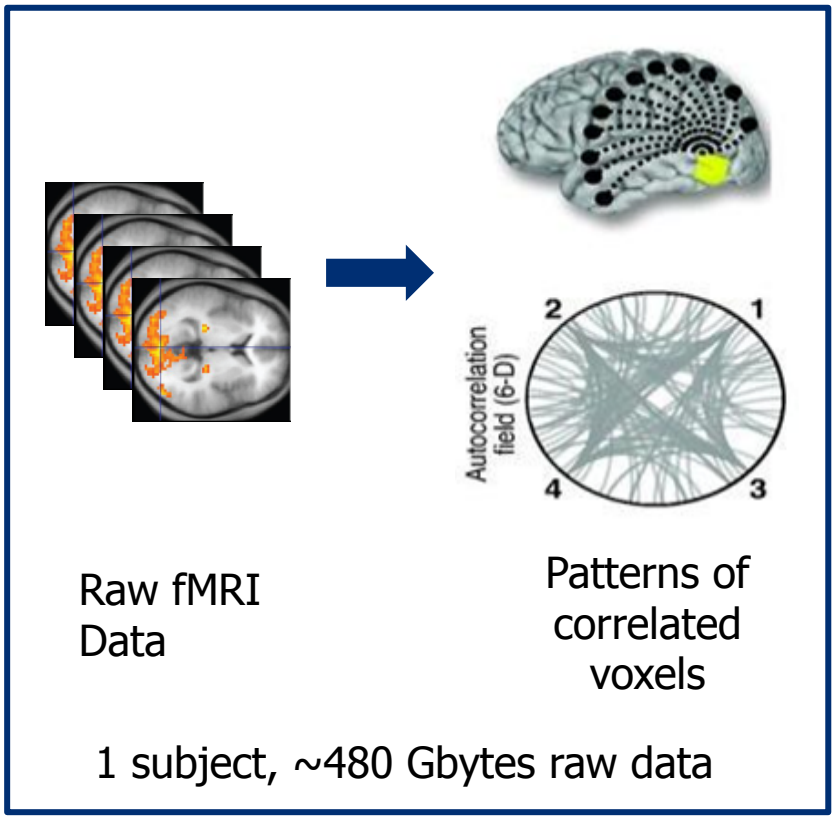


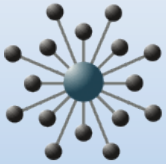
Image Sources: Princeton Neuroscience Institute and Wikipedia

Compute on Data where it is stored …
move queries, not the data

Lesson 4: One size does not fit all

- The application/source dictates the structure of the data:
  - Arrays
  - Tables
  - Documents
  - Other?
- Different applications mean different structure.
- A mismatch between Data and its Data store means the queries don't fit the data … leading to inefficiency or messy programming.

The structure of the data must match the Data store



ISTC
BIG DATA

So what will we be doing at our ISTC?

ISTC Research themes

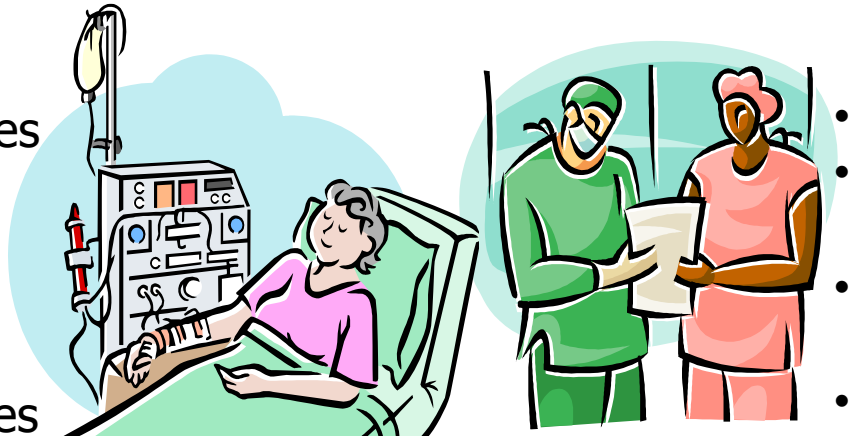
- Data Analytics & Processing Platforms
- Scalable Math and Algorithms
- Visualization
- Architecture

intel
MIT
BROWN
Carnegie Mellon University
W
UT
Portland State UNIVERSITY

Big Data in the real world

- Consider patient data in an Intensive Care Unit (e.g. MIMIC II data set*)

- EKG traces
- Blood oxygen
- Blood pressure
- EEG traces



- Demographic
- Caregiver notes
- Medical charts
- Lab test results
- Xray, MRI, etc.

The challenge … apply predictive analytics across all data … so we can show up to restart a heart before it stops beating!!!

* MIMIC: Multiparameter Intelligent Monitoring in Intensive Care, http://www.physionet.org/mimic2/

Big Data in the real world

Messy, heterogeneous, complex, streaming …

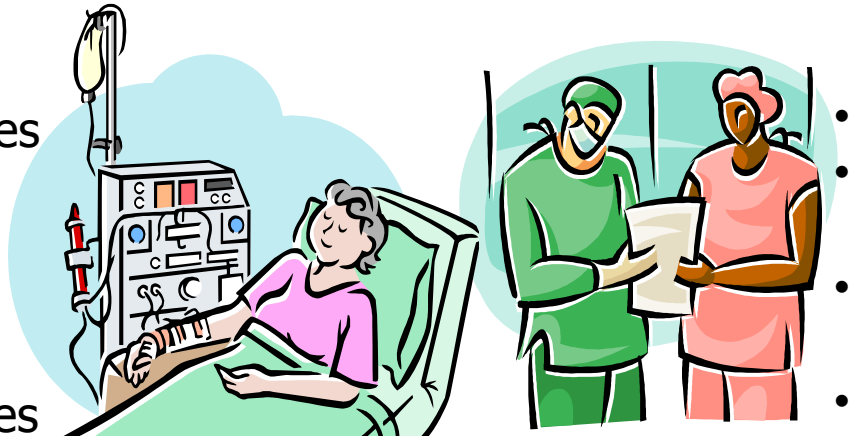
- Consider patient data in an Intensive Care Unit (e.g. MIMIC II data set*)

Arrays
- EKG traces

Time Series
- Blood oxygen

Time Series
- Blood pressure

Arrays
- EEG traces



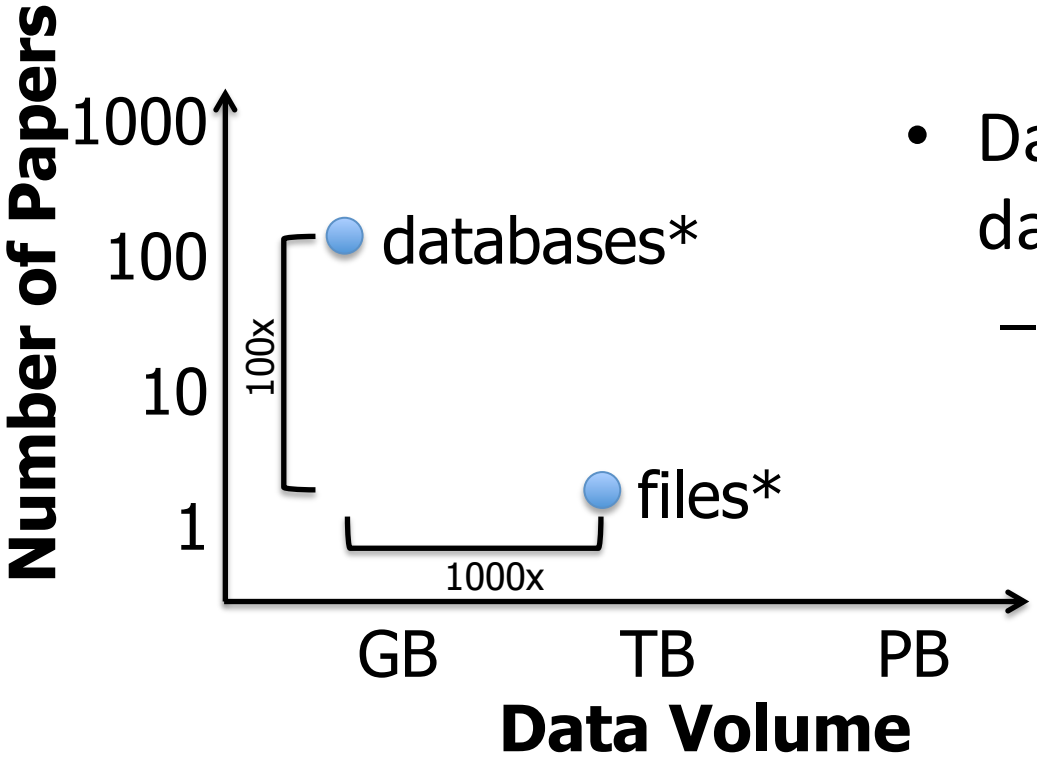
- Demographic — tables
- Caregiver notes — documents
- Medical charts — tables
- Lab test results — tables
- Xray, MRI, etc. — #images

Time series and tabular data are stored in a DBMS.
Other data? Flat files

* MIMIC: Multiparameter Intelligent Monitoring in Intensive Care, http://www.physionet.org/mimic2/

MIMC doesn't include images. We are talking to several groups to add an image database to our project

Analysis of published MIMICII papers



- Data in databases is used; data in files is not
  - Data in files is nearly equivalent to deleting the data

A disruptive idea: Match data to the data-store technology but present as a single Data Base Management system to the end-users … A disruptive idea we call *Polystore*.

Source: Vijay Gadepally of MIT Lincoln labs

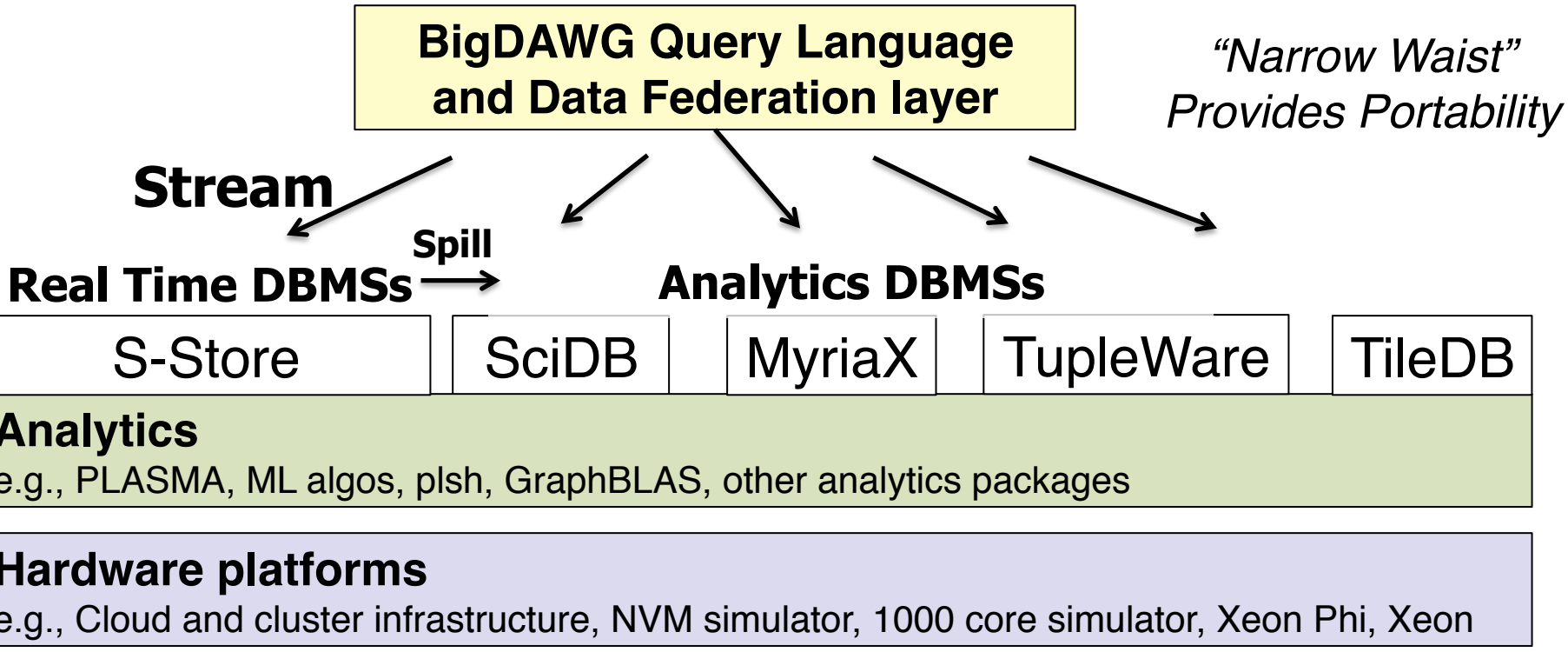
*Based on PhysioNet MIMIC2 ICU data

BigDawg: An integrated ***polystore*** system

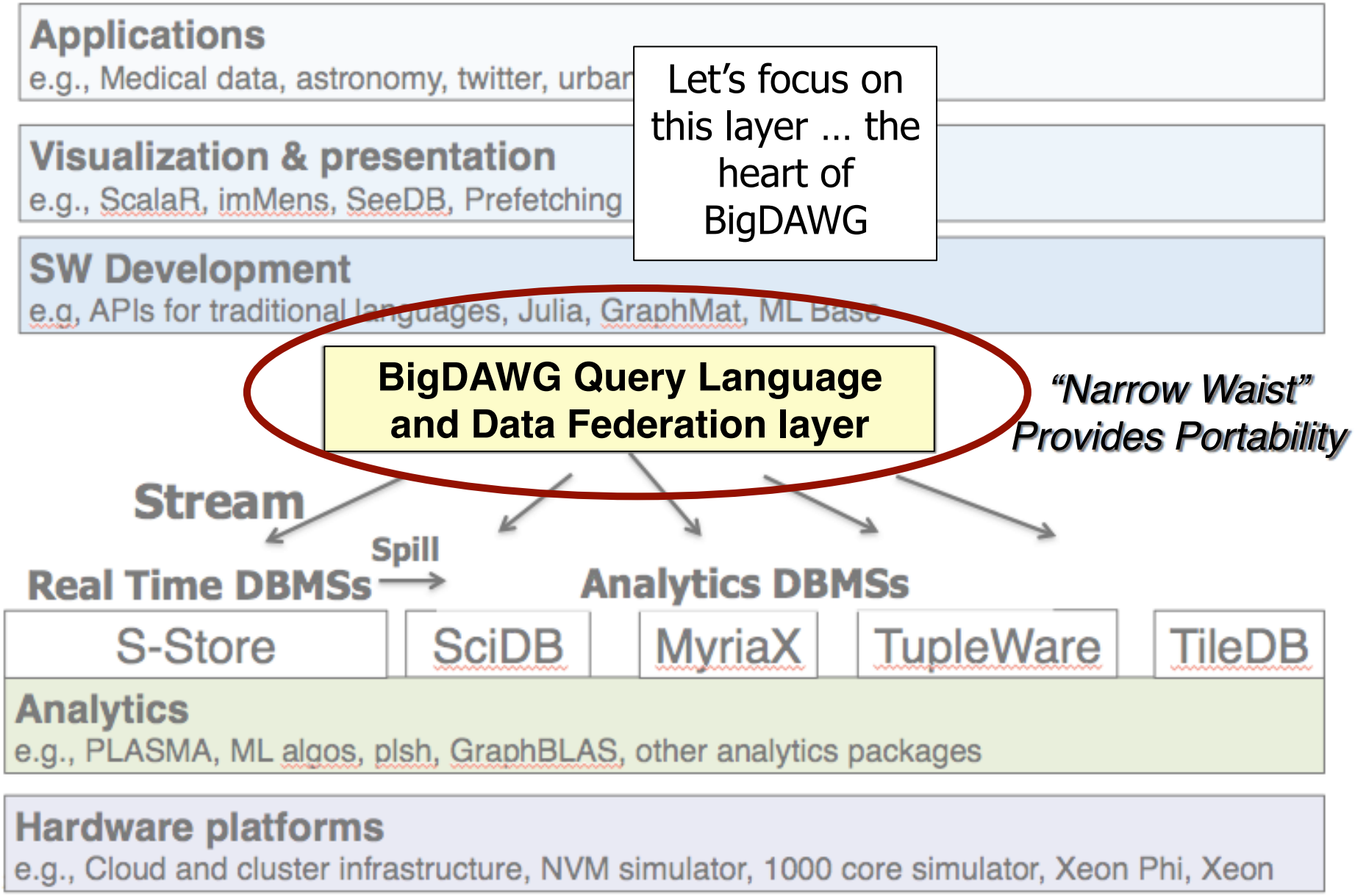
Applications
e.g., Medical data, astronomy, twitter, urban sensing, IoT

Visualization & presentation
e.g., ScalaR, imMens, SeeDB, Prefetching

SW Development
e.g, APIs for traditional languages, Julia, GraphMat, ML Base



BigDawg: An integrated ***polystore*** system

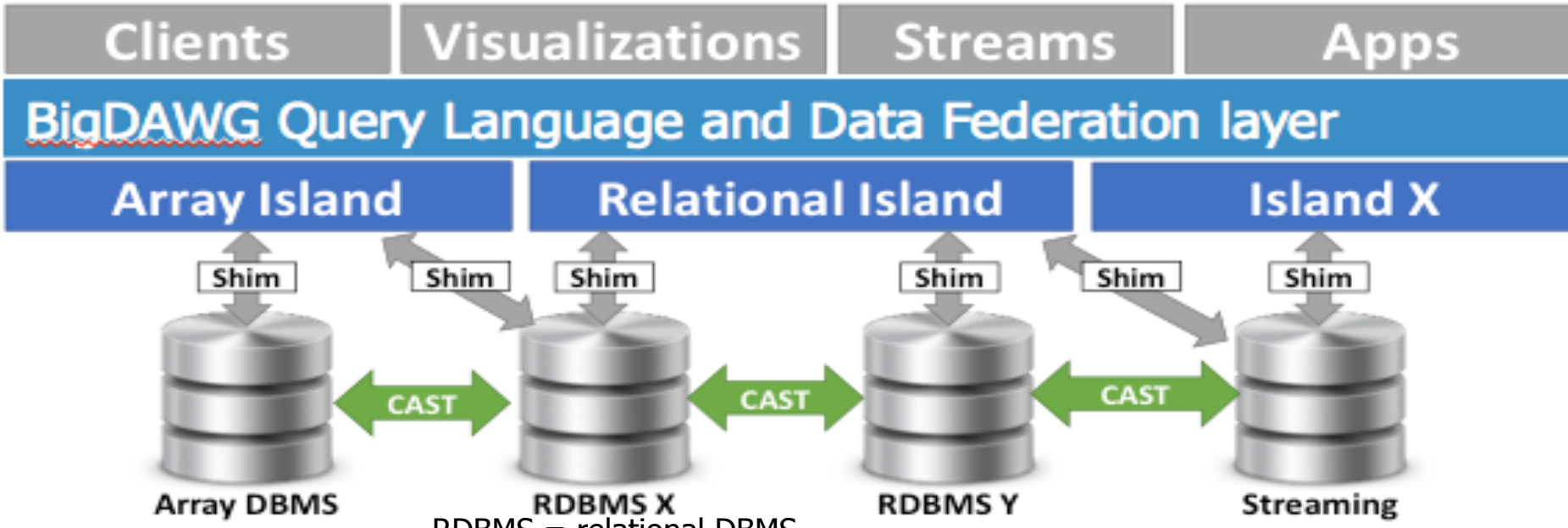


BigDAWG Data Federation

- Two Key Components:
  - BigDAWG Query Language or BQL:
    - the quest for “one query language to rule them all”
  - BigDAWG Data Federation API:
    - **Islands**: a collection of data stores that share a data model and query language
    - **Shims**: to translate queries between islands
    - **Casts**: to move data from one island to another

High risk transformative research … many people think this is impossible.

Based on ISTC research over the last 3 years, we think we know how to do this



RDBMS = relational DBMS

Our VLDB'2015 Demo

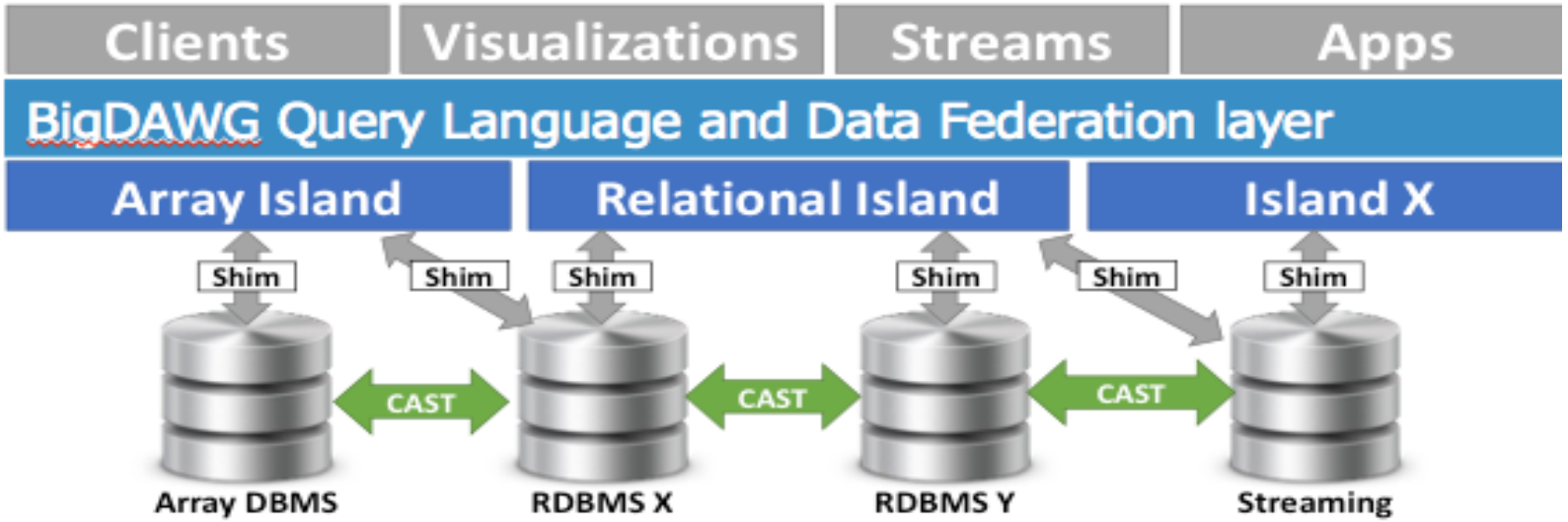
A Demonstration of the BigDAWG Multi-Database System

A. Elmore	J. Duggan	M. Stonebraker	M. Balazinska	U. Cetintemel	V. Gadepally	J. Heer	B. Howe	J. Kepner
Univ. of Chicago	Northwestern	MIT	Univ. of Wash.	Brown	MIT-LL	Univ. of Wash.	Univ. of Wash.	MIT-LL
T. Kraska	S. Madden	D. Maier	T. Mattson	S. Papadopoulis	J. Parkhurst	N. Tatbul	M. Vartek	S. Zdonik
Brown	MIT	Portland St U.	Intel	Intel / MIT	Intel	Intel / MIT	MIT	Brown

ABSTRACT

This paper presents BigDAWG, a reference implementation of a new architecture for "Big Data" applications. Such applications

BigDAWG stores MIMIC II in a mixture of backends, including Postgres, which stores the patient metadata, SciDB [4], which stores the historical waveform data in a time-series (array) database,



Our VLDB'2015 Demo

A Demonstration of the BigDAWG Multi-Database System

A. Elmore, Univ. of Chicago; J. Duggan, Northwestern; M. Stonebraker, MIT; M. Balazinska, Univ. of Wash.; U. Cetintemel, Brown; V. Gadepally, MIT-LL; J. Heer, Univ. of Wash.; B. Howe, Univ. of Wash.; J. Kepner, MIT-LL

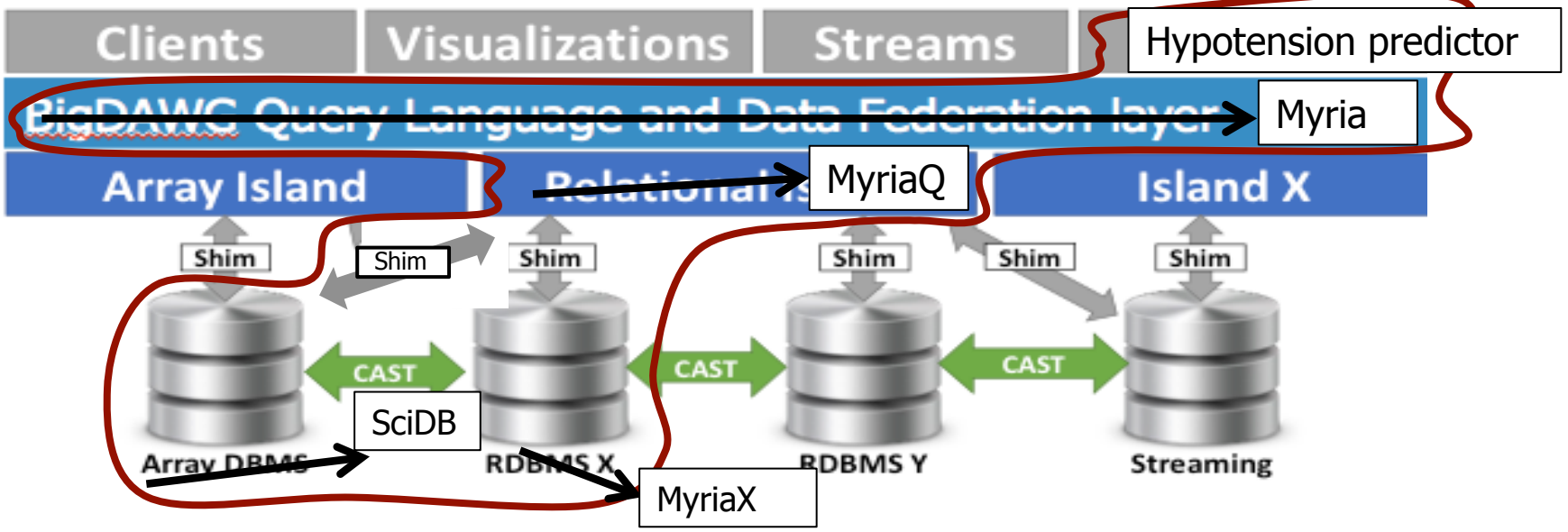
T. Kraska, Brown; S. Madden, MIT; D. Maier, Portland St U.; T. Mattson, Intel; S. Papadopoulis, Intel / MIT; J. Parkhurst, Intel; N. Tatbul, Intel / MIT; M. Vartek, MIT; S. Zdonik, Brown

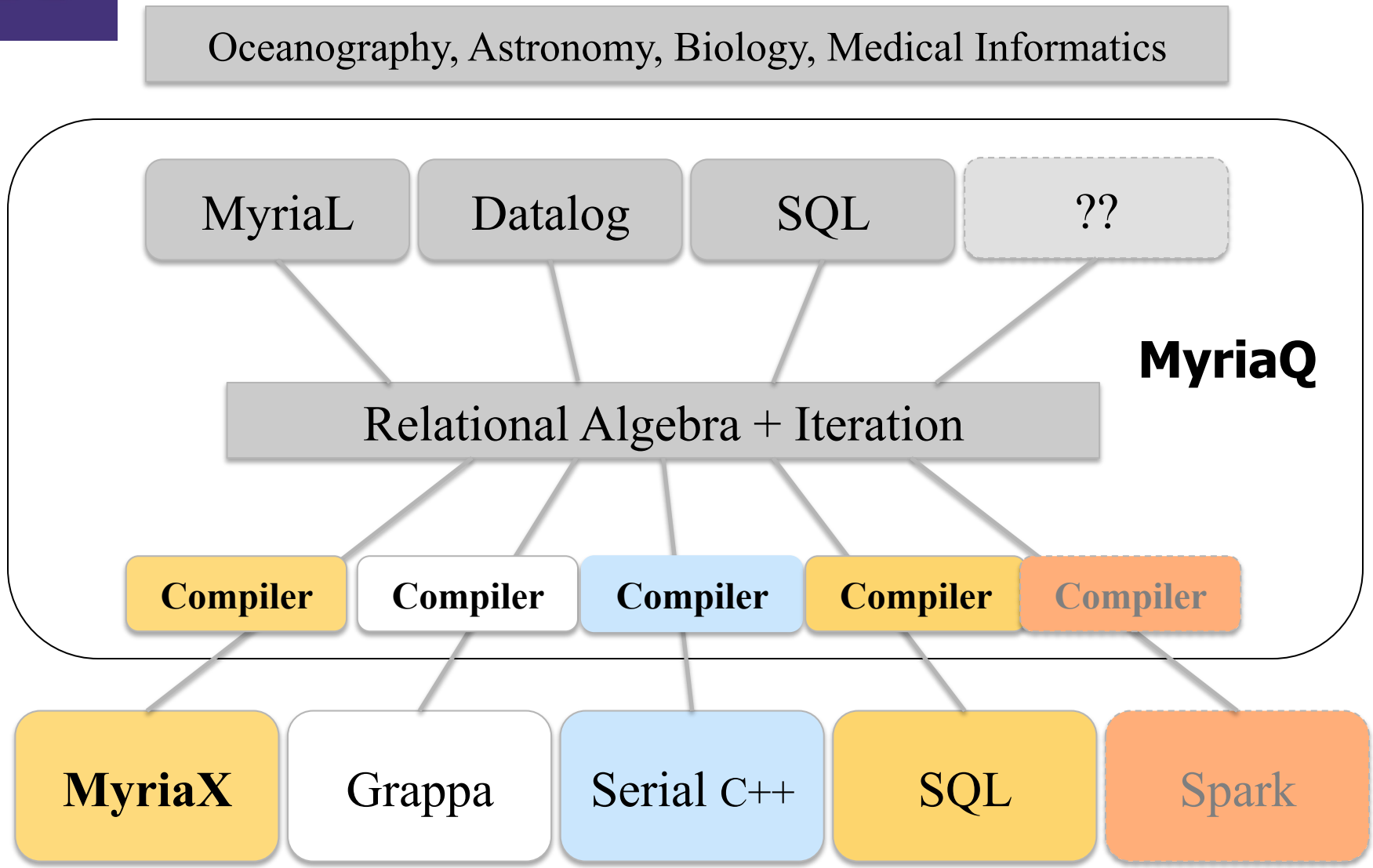
We have preliminary results targeting MyriaX, SciDB, S-store, D4M, graphulo and several visualization packages.

AB

This paper presents BigDAWG, a reference implementation of a new architecture for "Big Data" applications. Such applications

including Postgres, which stores the patient metadata, SciDB [4], which stores the historical waveform data in a time-series (array) database,

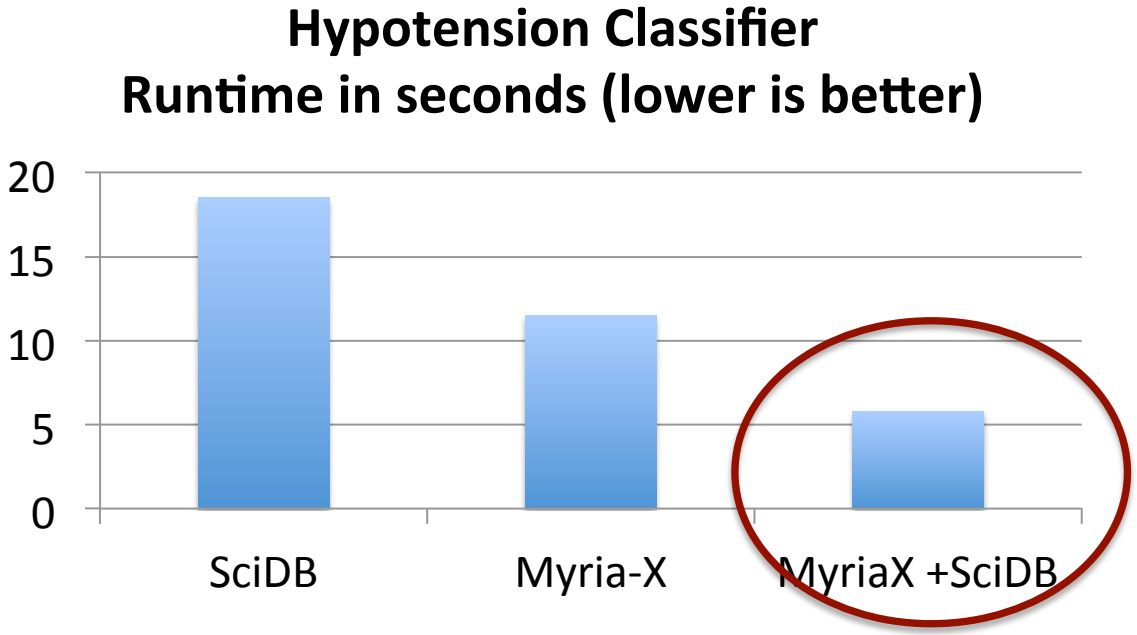




Magda Balazinsk, Bill Howe, Dan Suciu, Dan Halperin http://myria.cs.washington.edu/

The App: Hypotension Predictor

- Problem: blood pressure drops (hypotension) →shock → death. Early intervention is key for survival.
- Solution: Machine learning over heterogeneous data (from MIMIC II) to identify patients about to suffer from a severe drop in blood pressure?
- Algorithm (from Saeed and Mark*) build a classifier .. Haar transforms over MIMICII time series data, summarize as histograms, and performs a K nearest neighbor search. Correlate with patient data.



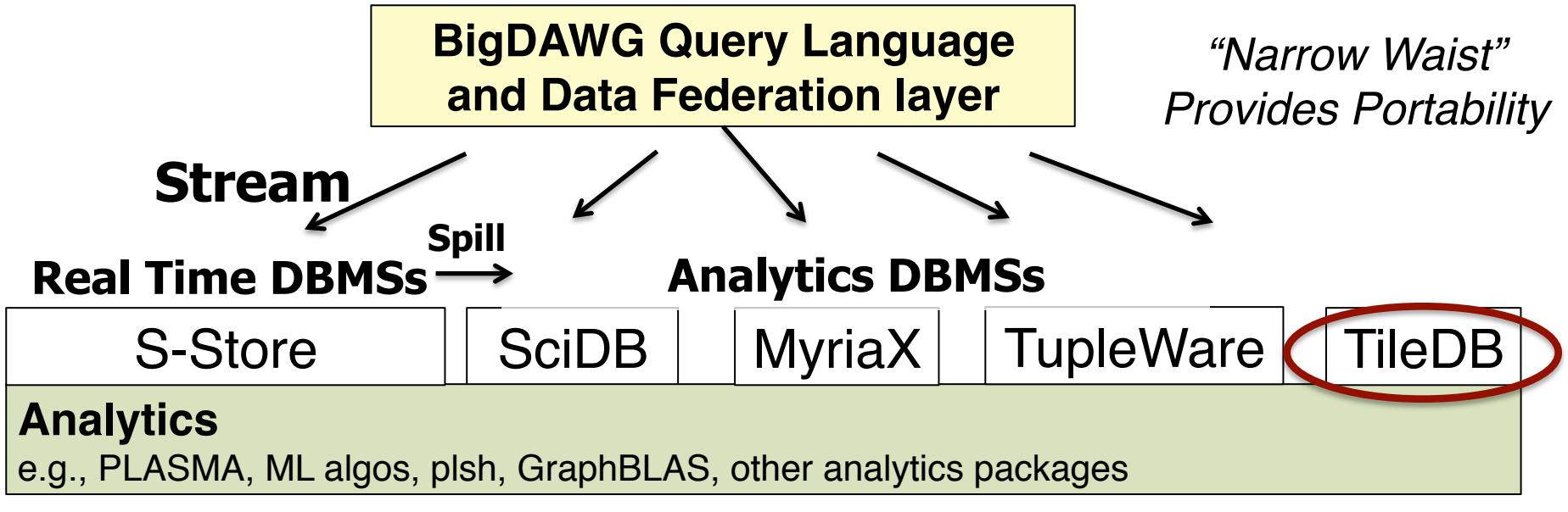
Source: Magdalena Balazinska and Brandon Haynes, university of Washington.
*A Novel Method for the Efficient Retrieval of Similar Multiparameter Physiologic Time Series Using Wavelet-Based Symbolic Representations. Mohammed Saeed and Roger Mark, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839671/

BigDawg: An integrated ***polystore*** system

Applications
e.g., Medical data, astronomy, twitter, urban sensing, IoT

Visualization & presentation
e.g., ScalaR, imMens, SeeDB, Prefetching

SW Development
e.g, APIs for traditional languages, Julia, GraphMat, ML Base

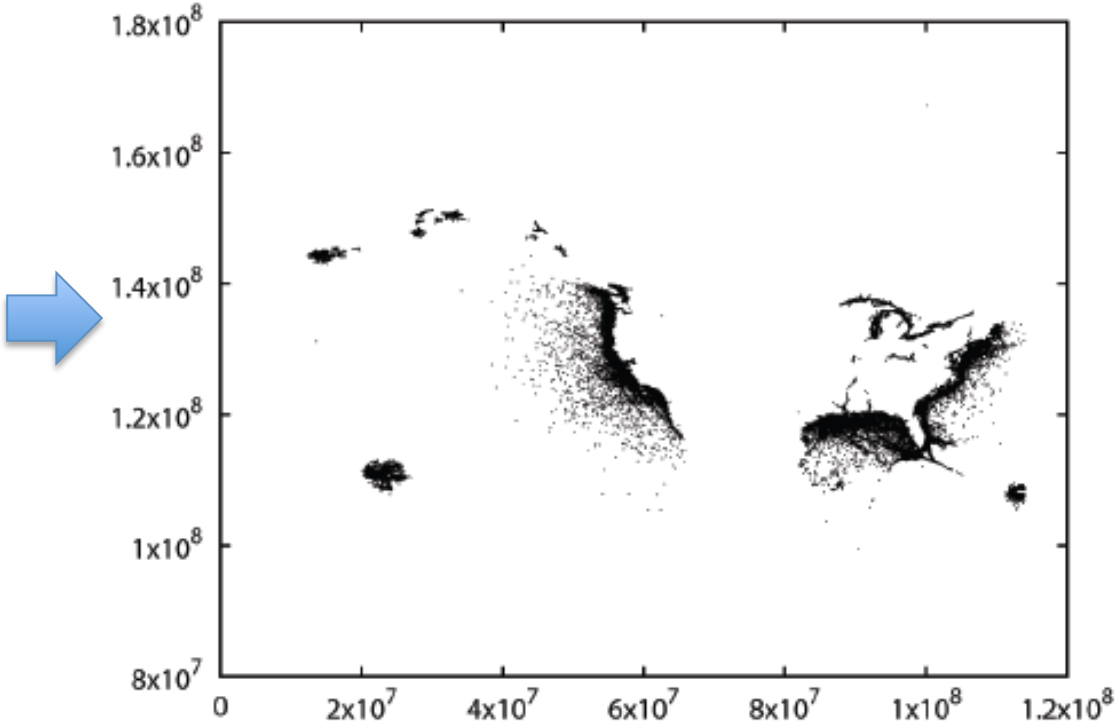


Hardware platforms
e.g., Cloud and cluster infrastructure, NVM simulator, 1000 core simulator, Xeon Phi, Xeon

Arrays in Big Data problems

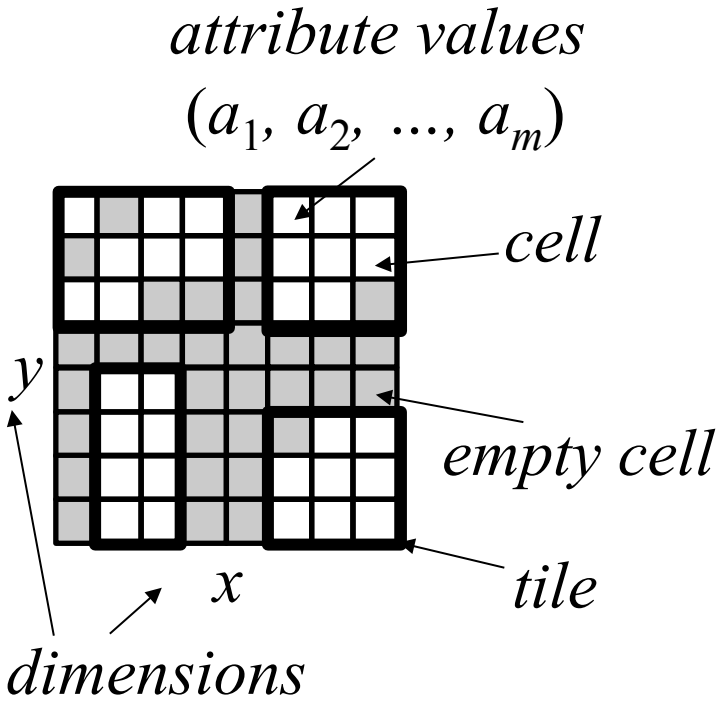
- Data is often naturally considered as an array:
  - An object with multiple dimensions (e.g. 2)
  - The dimensions define a logical coordinate space
  - A cell "exists" at each point in the coordinate space.
  - A cell has one or more attributes which collectively define the "value" at that cell.

- Data is usually sparse
  - E.G. the AIS data set showing ship locations as a function of time in and around U.S. waters

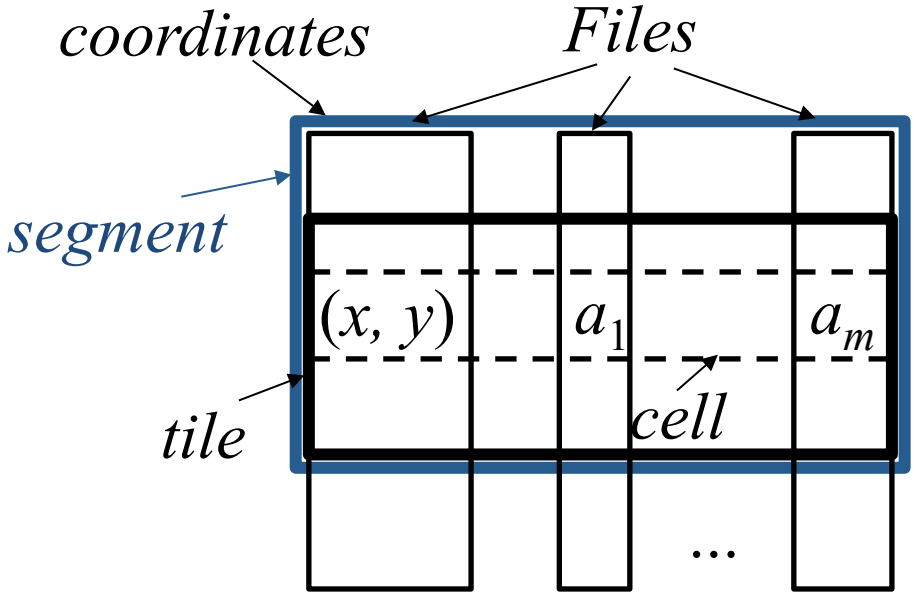


TileDB a new array data storage manager: optimized for Sparse Arrays

Logical representation



Physical representation



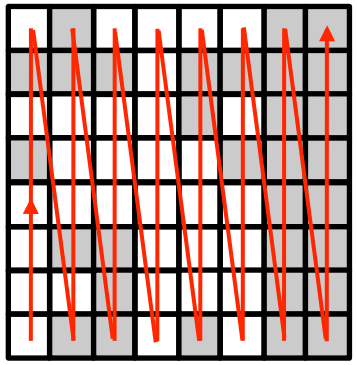
Tile: *Atomic unit of processing*

Segment: *Atomic unit of I/O*

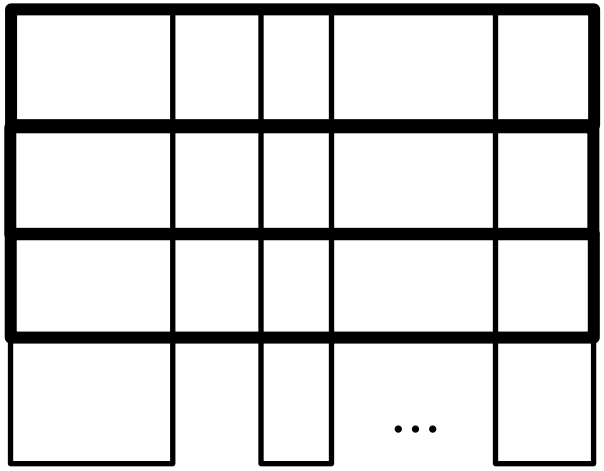
Manage array storage as tiles of different shape/size in the index space, but with ~equal number of non-empty cells

Loading data into TileDB

Impose a **global order**

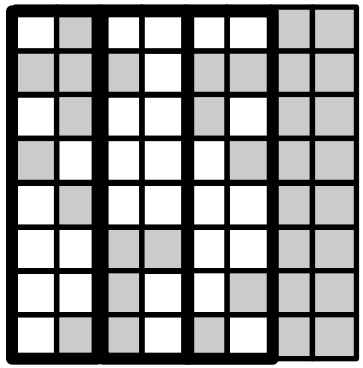


Files sorted on cell id



sort

Pack cells, create book-keeping structures (`make_tiles`)



Logical representation

TileDB use case: genomic data

- Consider data representing differences from a reference genome stored in the standard "genomic variant Call format" (gVCF).
  - A **sample** is one subject's exome (the portion of the genome that is "translated" into protein) … ~10MB
  - Each line in the gVCF file corresponds to a **range** of chromosome positions.
  - The value at each position is a measure of the probabilistic similarity/ dissimilarity to a reference sample at that position.

gVCF example segment

```
chr20  287125  .  T  .  .  PASS  END=287136;BLOCKAVG_min30p3a  GT:DP:GQX:MQ  0/0:40:78:40
chr20  287137  .  G  .  .  LowGQX  .  GT:DP:GQX:MQ  0/0:42:11:42
chr20  287138  .  C  .  .  PASS  END=287178;BLOCKAVG_min30p3a  GT:DP:GQX:MQ  0/0:36:96:42
chr20  287179  .  C  T  310.01  PASS  BaseQRankSum=-0.721;DP=37;Dels=0.00;FS=14.994;HaplotypeScore=0.0000;MLEAC=1;MLEAF=0.500;MQ=52.29;MQ0=
0;MQRankSum=-1.091;QD=8.38;ReadPosRankSum=-1.963;SB=-1.901e+01  GT:AD:DP:GQ:PL:MQ:GQX  0/1:24,13:37:99:340,0,810:52:99
chr20  287180  .  G  .  .  PASS  END=287245;BLOCKAVG_min30p3a  GT:DP:GQX:MQ  0/0:32:78:49
chr20  287246  .  G  A  567.01  PASS  BaseQRankSum=-0.718;DP=33;Dels=0.00;FS=5.093;HaplotypeScore=3.2995;MLEAC=1;MLEAF=0.500;MQ=49.01;MQ0=0
;MQRankSum=1.050;QD=17.18;ReadPosRankSum=0.129;SB=-2.920e+02  GT:AD:DP:GQ:PL:MQ:GQX  0/1:13,20:33:99:597,0,343:49:99
chr20  287247  .  C  .  .  PASS  END=287259;BLOCKAVG_min30p3a  GT:DP:GQX:MQ  0/0:27:75:46
```

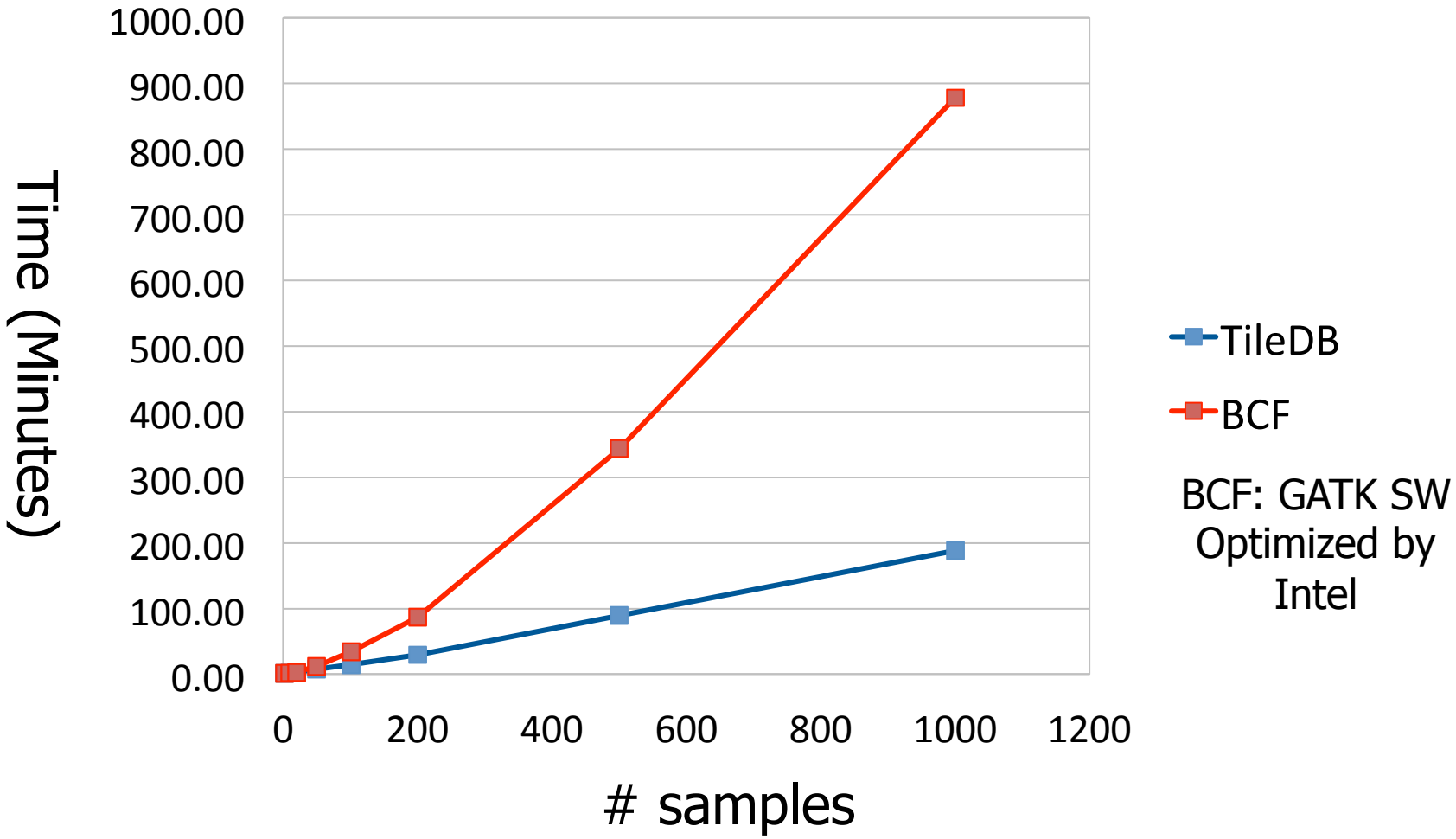
A genomics center has lots of these files … E.G. the Broad has 1K in their public data set and 100K in their internal data sets! These numbers will grow radically over time.

Processing over gVCF data

- An important use of gVCF data is compare many genomes to identify mutations at specific positions (e.g. joint genotyping).
- Our collaborators at the Broad Genomics institute* gave us a joint genotyping proxy application:
  - Load gVCF files and compute median values over 5000 positions across all samples

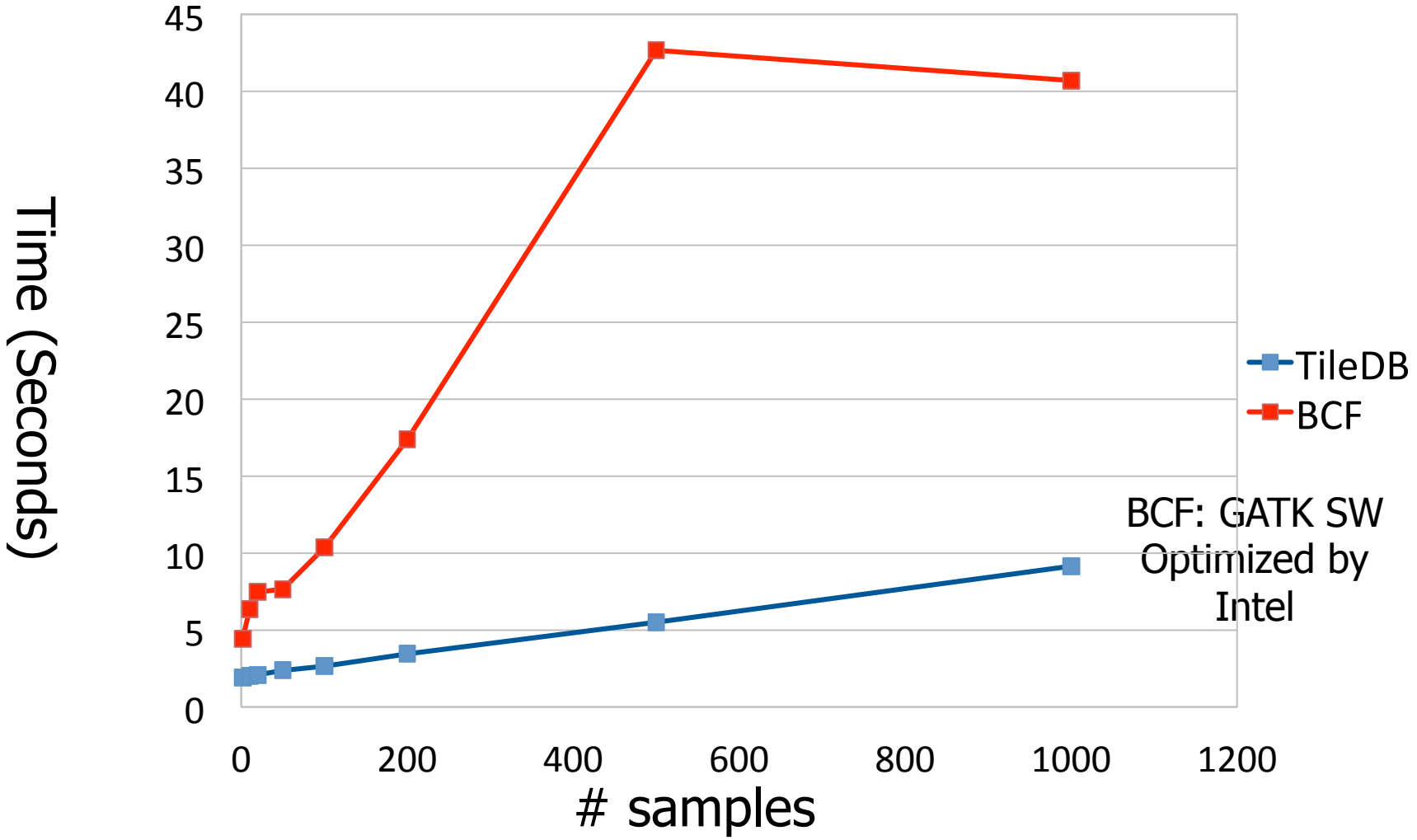
* The authors of the famous GATK genome processing software.

Data Loading cost



Intel® Xeon® E5 2697 v2 CPU, 12 cores, dual socket, 128 GB RAM, CentOS6.6, Western Digital 4 TB WD4000F9YZ-0 as a ZFS RAID0 pool. Single thread/core results.

Joint Genotyping Benchmark*



*Benchmark jointly developed by Intel and the Broad Genomics Institute. Each sample is 10MB. Compute correlations across samples at ~5000 positions.

Intel® Xeon® E5 2697 v2 CPU, 12 cores, dual socket, 128 GB RAM, CentOS6.6, Western Digital 4 TB WD4000F9YZ-0 as a ZFS RAID0 pool. Single thread/core results.

BigDawg: An integrated ***polystore*** system

Applications
e.g., Medical data, astronomy, twitter, urban sensing, IoT

Visualization & presentation
e.g., ScalaR, imMens, SeeDB, Prefetching

SW Development
e.g, APIs for traditional languages, Julia, GraphMat, ML Base

BigDAWG Query Language and Data Federation layer

"Narrow Waist" Provides Portability

Stream

Spill

Real Time DBMSs

Analytics DBMSs

S-Store | SciDB | MyriaX | TupleWare | TileDB

Analytics
e.g., PLASMA, ML algos, plsh, GraphBLAS, other analytics packages

Hardware platforms
e.g., Cloud and cluster infrastructure, NVM simulator, 1000 core simulator, Xeon Phi, Xeon

Visualization

- Visualization is the primary way user's consume data
- Has been an afterthought in data-intensive systems
- Tremendous opportunities to improve performance and usability of vis by tightly coupling with data processor

SeeDB

Parameswaran (Stanford) et. al. VLDB 2014

- Key idea: help users find interesting things in data sets
- Visualizations are (usually) plots of 2 (or 3) attributes
  - E.g., sales by region
- What makes a visualization interesting?
  - Relevant to user
  - Highly variable
  - Not too many or too few distinct values

Approach

- **SeeDB** searches through all visualizations of 2,3,…,n dimensions
  - Using an efficient techniques to compute *all* group by queries
- Finds most *interesting* view

Given query Q over subset of data D' of database D

interestingness = Q(D') – Q(D-D')

One of many possible metrics

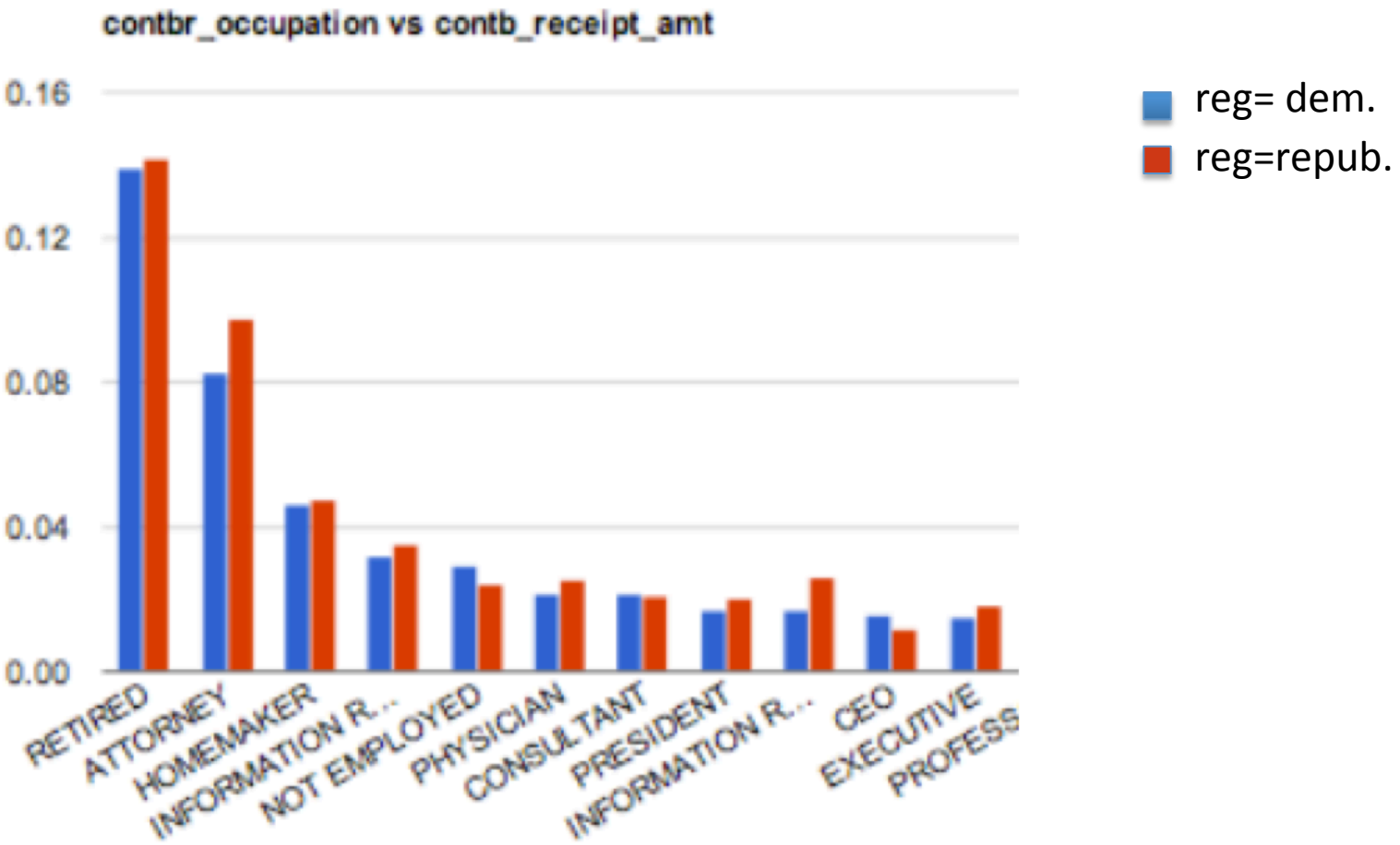
i.e., data with most variation in user-specified D' vs entire D

Example: *Compute average sales-by-region (Q) in database D, where user focuses on subset of sales of electronics (D')*

Max(Q(D') – Q(D-D')) = places where sales-by-region of electronics most different from sales-by-region of all other products

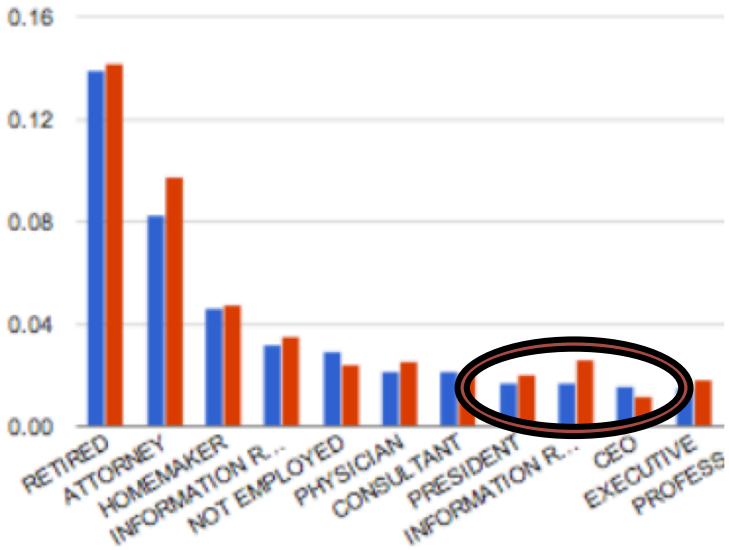
“Most Interesting” Group Bys

Contributor Profession vs Fraction $’s



Scorpion
(Sam Madden and Eugene Wu, MIT)

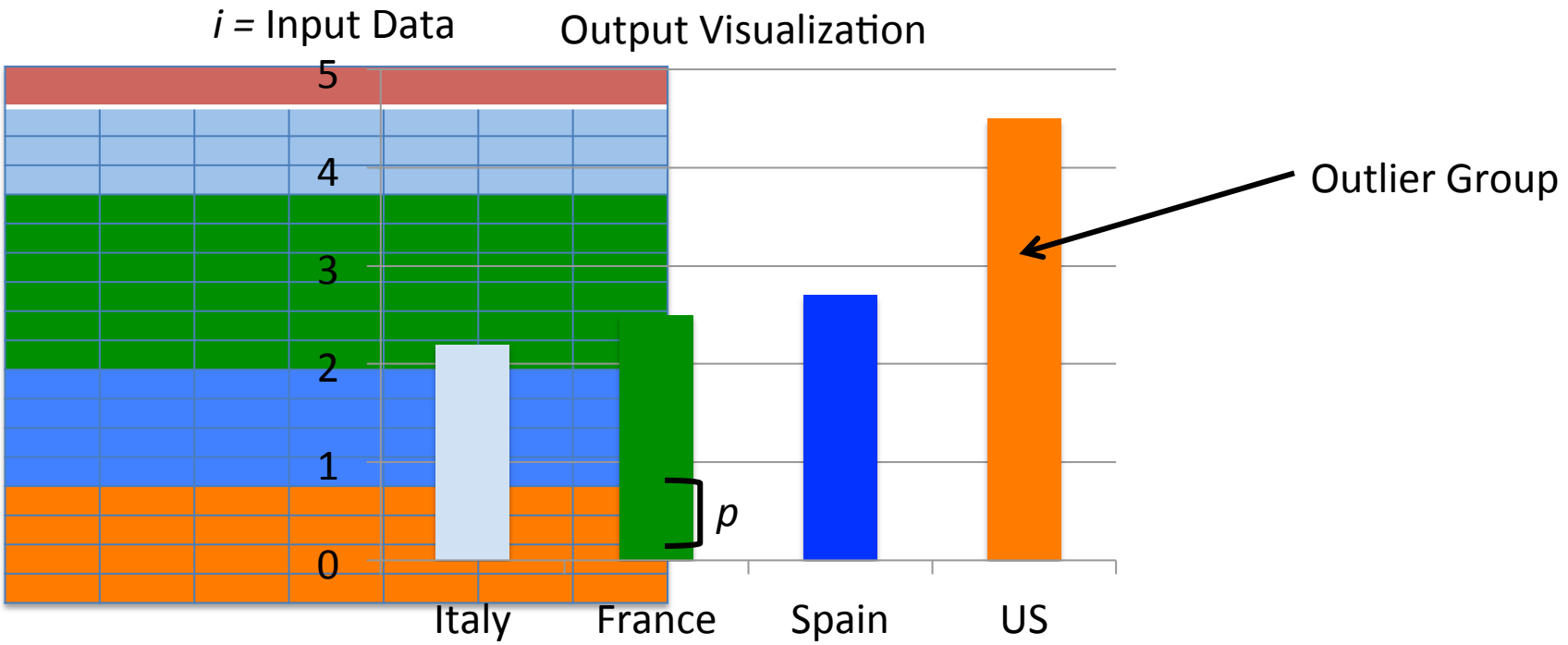
- After SeeDB: you found something interesting, now what?



- Common problem: outliers
- Need: a method to discover *why* outliers exist

Definition of Why

Given an outlier group, find a *predicate* over the inputs that makes the output no longer an outlier.

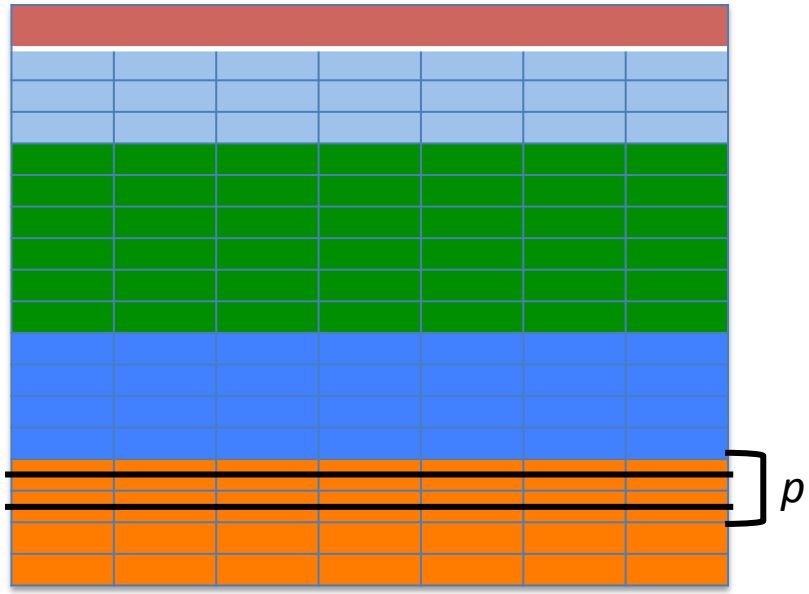


p = predicate

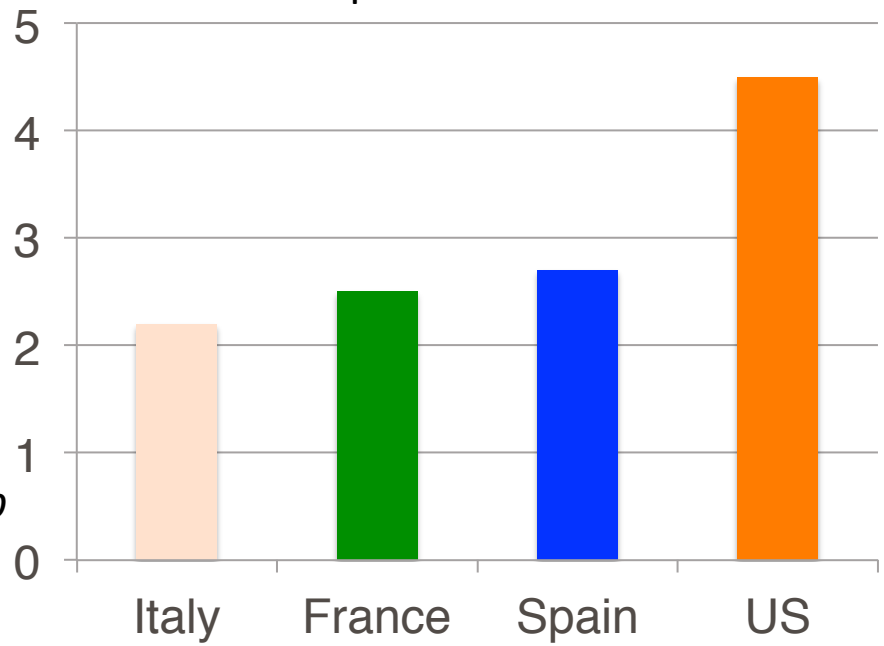
Definition of Why

Given an outlier group, find a *predicate* over the inputs that makes the output no longer an outlier.

i = Input Data



Output Visualization

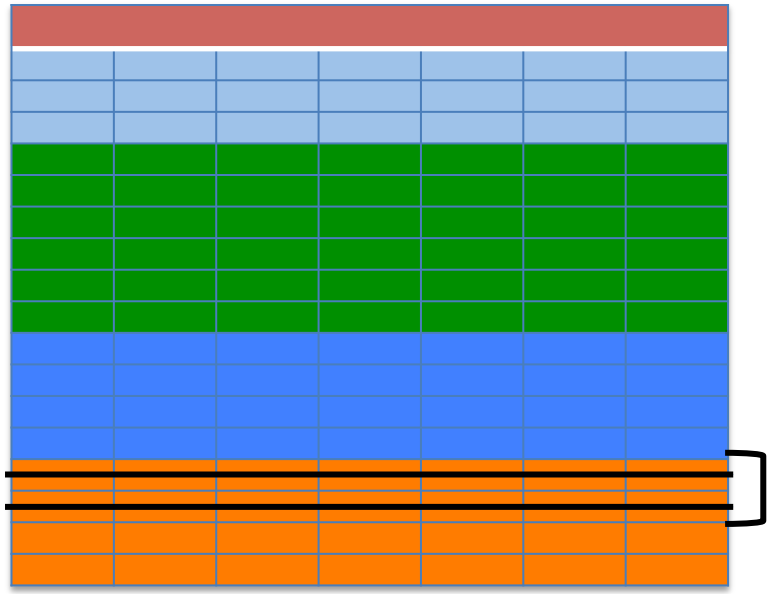


p = predicate

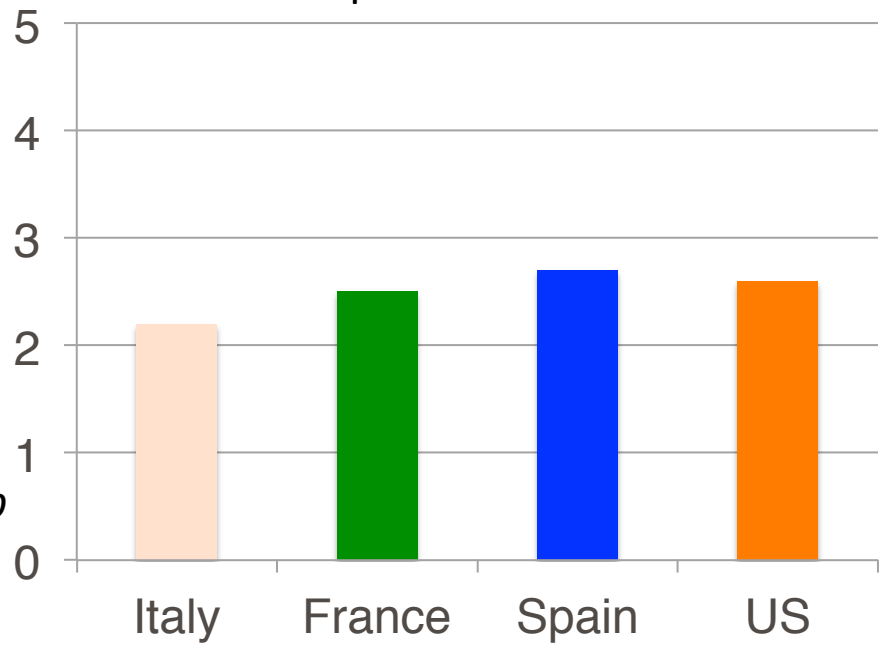
Definition of Why

Given an outlier group, find a *predicate* over the inputs that makes the output no longer an outlier.

i = Input Data



Output Visualization



Removing the predicate makes US no longer an outlier

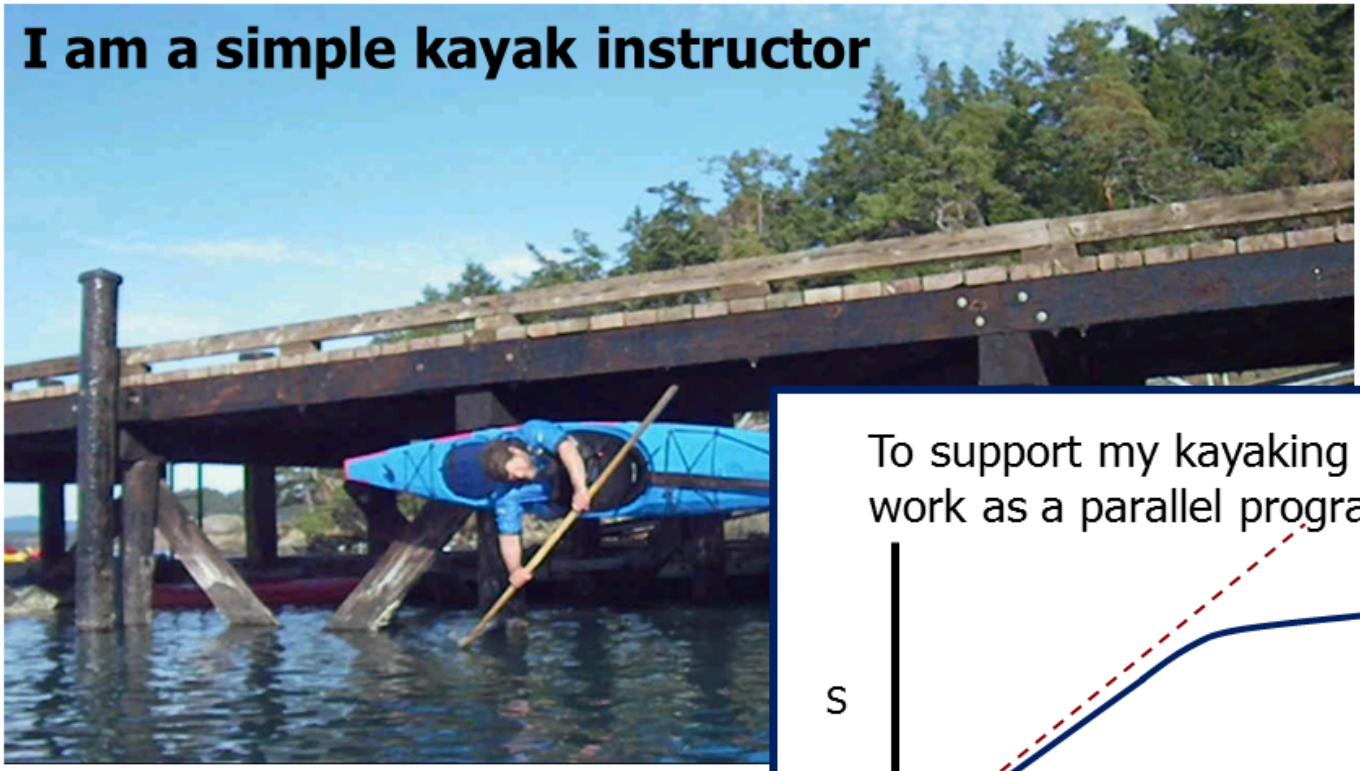
{Bill Gates, Steve Ballmer}

p: Company = MSFT

What are common properties of those records?

Am I ready to change my Introduction slide?

Introduction



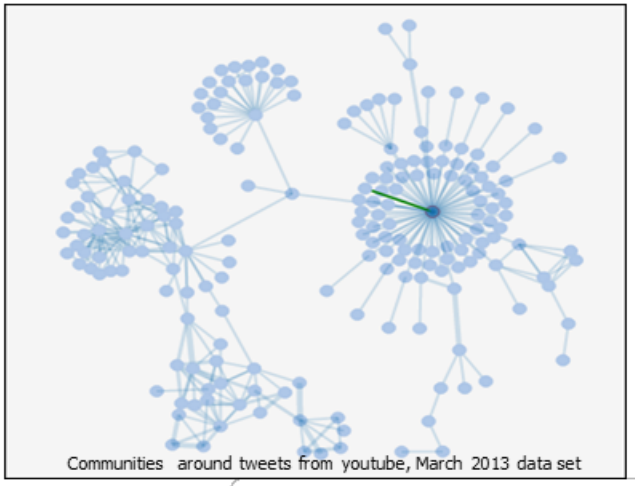
To support my kayaking habit I work as a parallel programmer

Which means I know how to turn math into lines on a speedup plot

Am I ready to change my Introduction slide?

Introduction

To support my kayaking habit



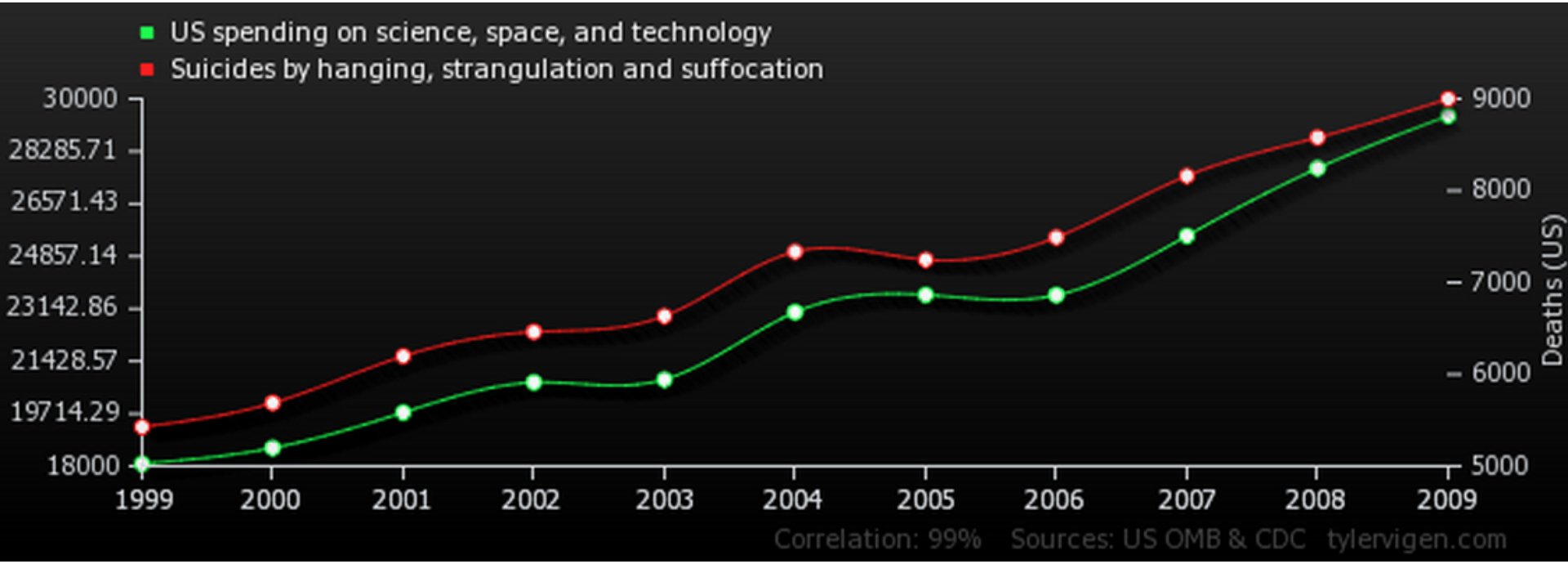
I create technologies that make people buy things they don't really need.

Big Concerns over Big Data

- Big Data methods find hidden information buried in massive piles of data
- If one is not careful:
  - Over fitting simple models … confuse your self with noise
  - Becoming confused by spurious correlations
  - Bad data can't magically give you good knowledge.
  - Pretending that prediction is knowledge.

Spurious Correlations

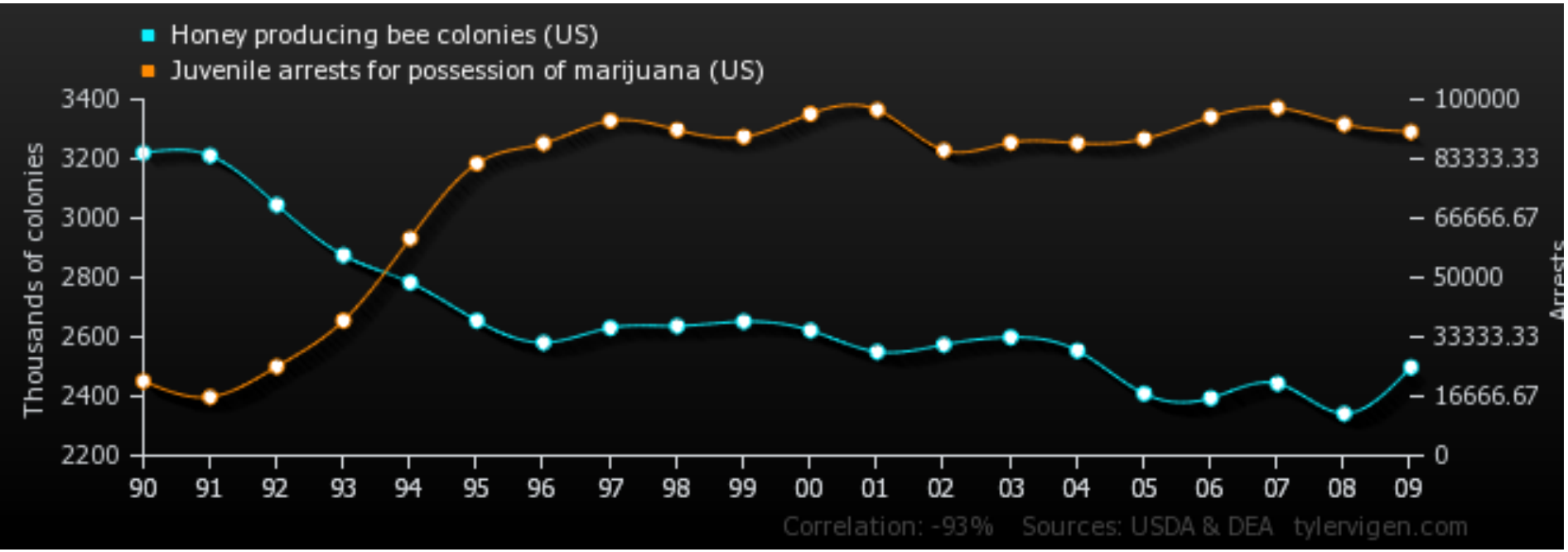
- Does correlation imply causality?



Source: http://www.tylervigen.com/, collected 10/15/2014

Spurious Correlations

- Does correlation imply causality?



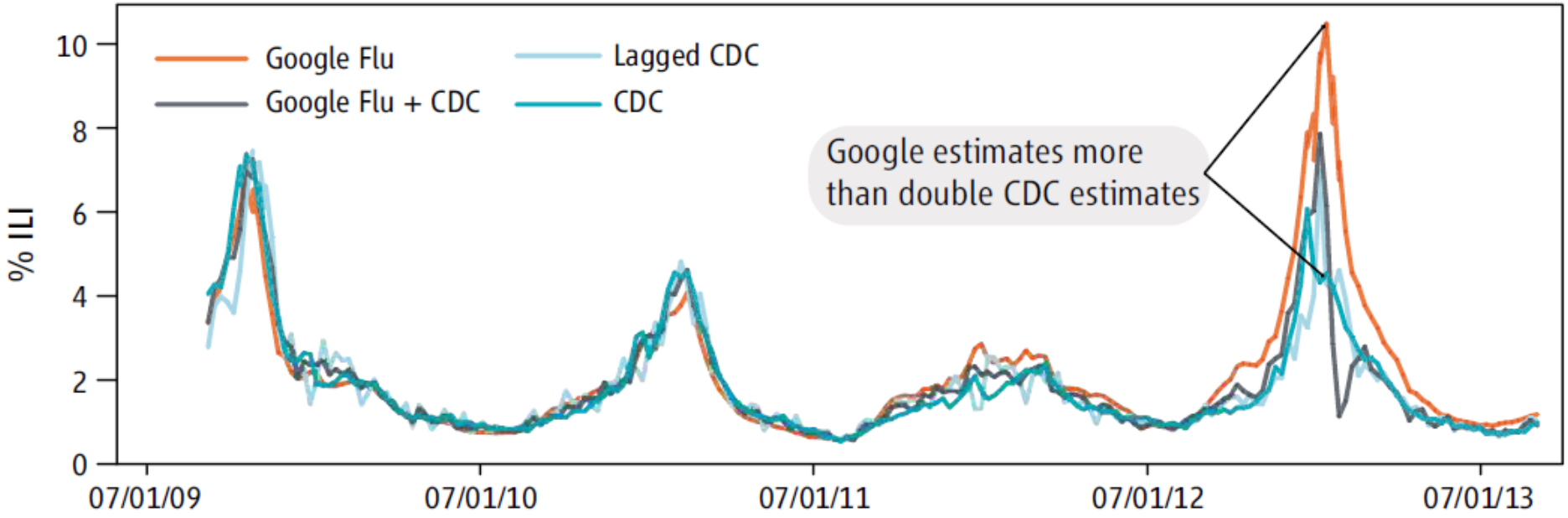
Source: http://www.tylervigen.com/, collected 10/15/2014

Google Flu

- Google claims that they can track influenza outbreaks in the U.S. just by tracking the queries that people post in their web searches.
- The concept is interesting, but there are issues.
- My biggest problem is that as of fall'Google has never published the details of which queries they track or how they process them to lead to their predictions.
  - It's hard to do science when there is no transparency

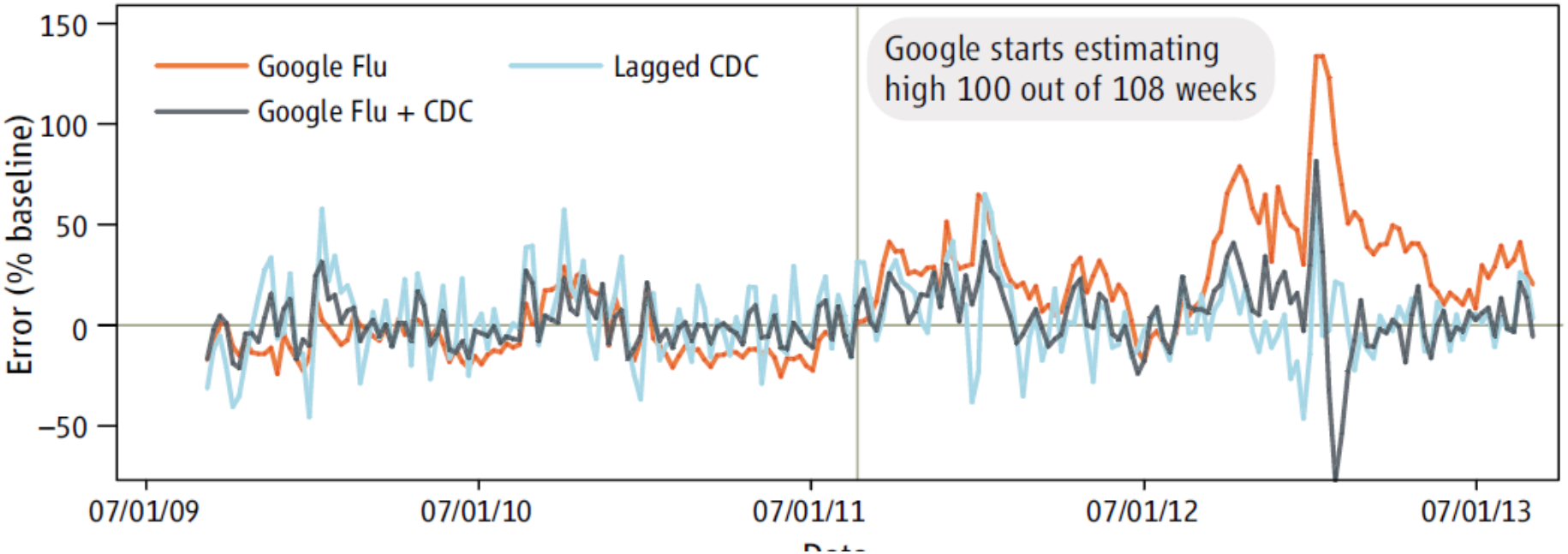
Google results are poor and misleading

- Compare Google Flu results to CDC's results (based on verified records doctor's visits.

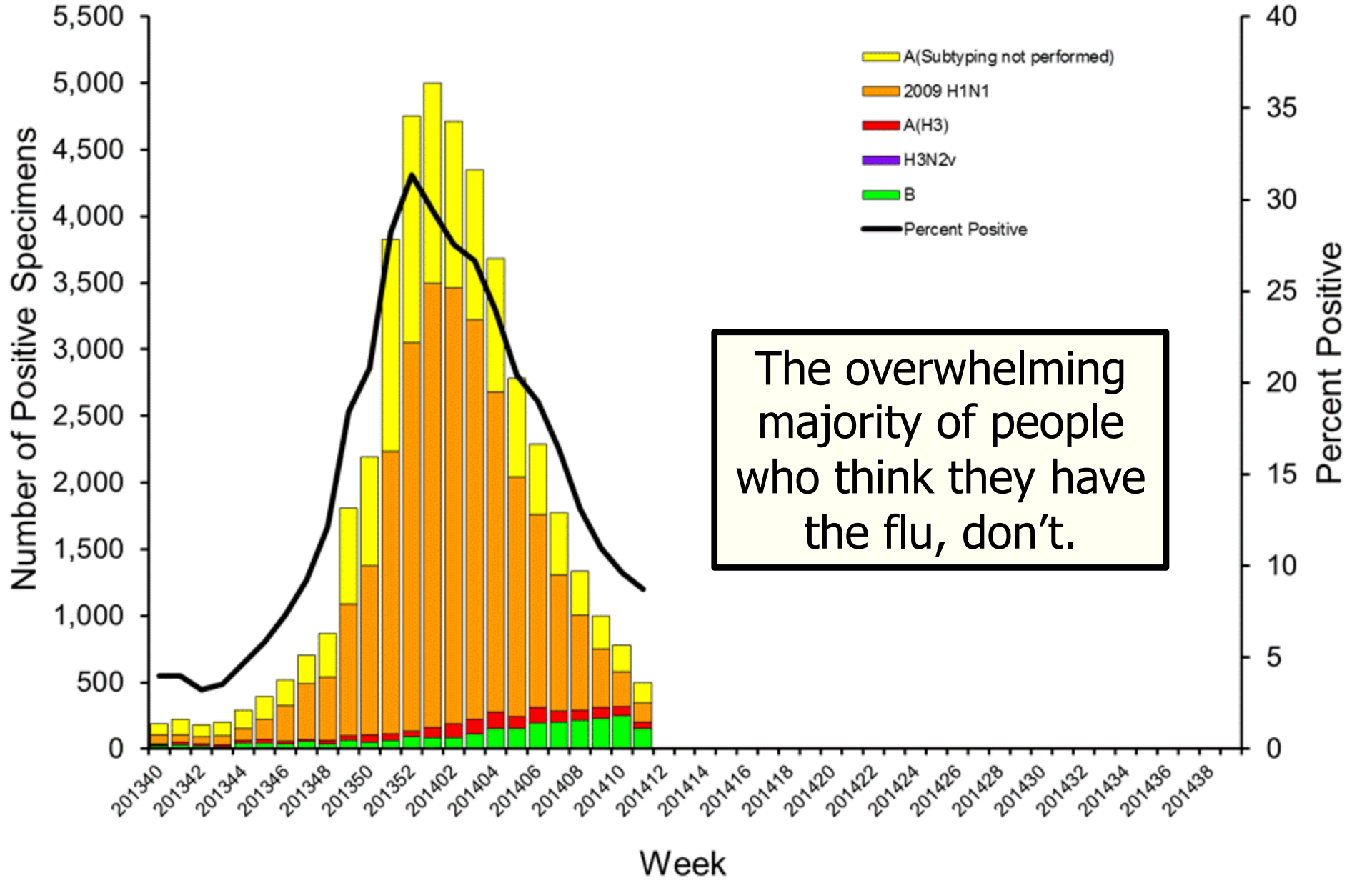


Google results are poor and misleading

- Compare Google Flu results to CDC's results (based on verified records doctor's visits.

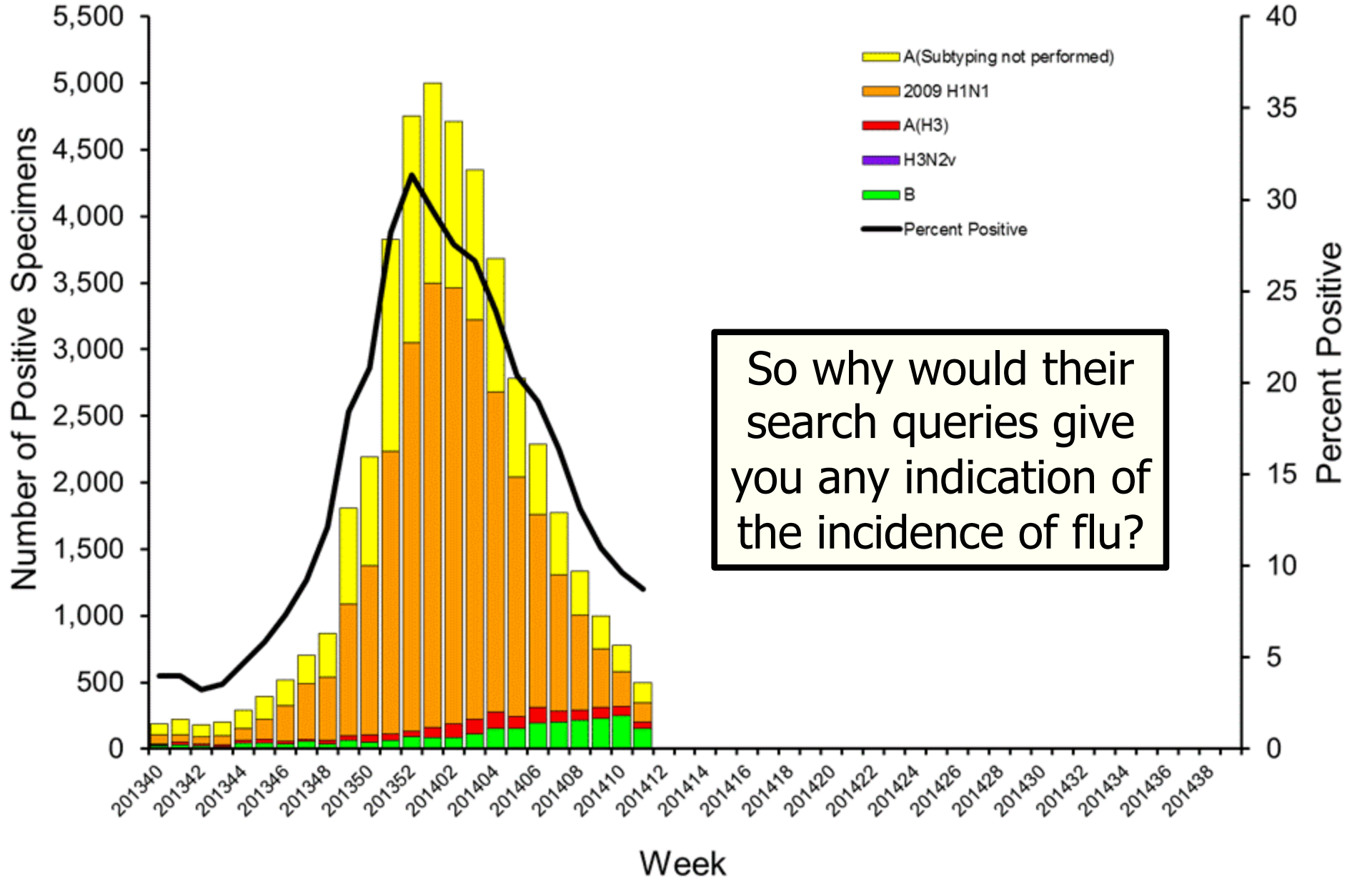


Influenza Positive Tests Reported to CDC by U.S. WHO/NREVSS Collaborating Laboratories, National Summary, 2013-14



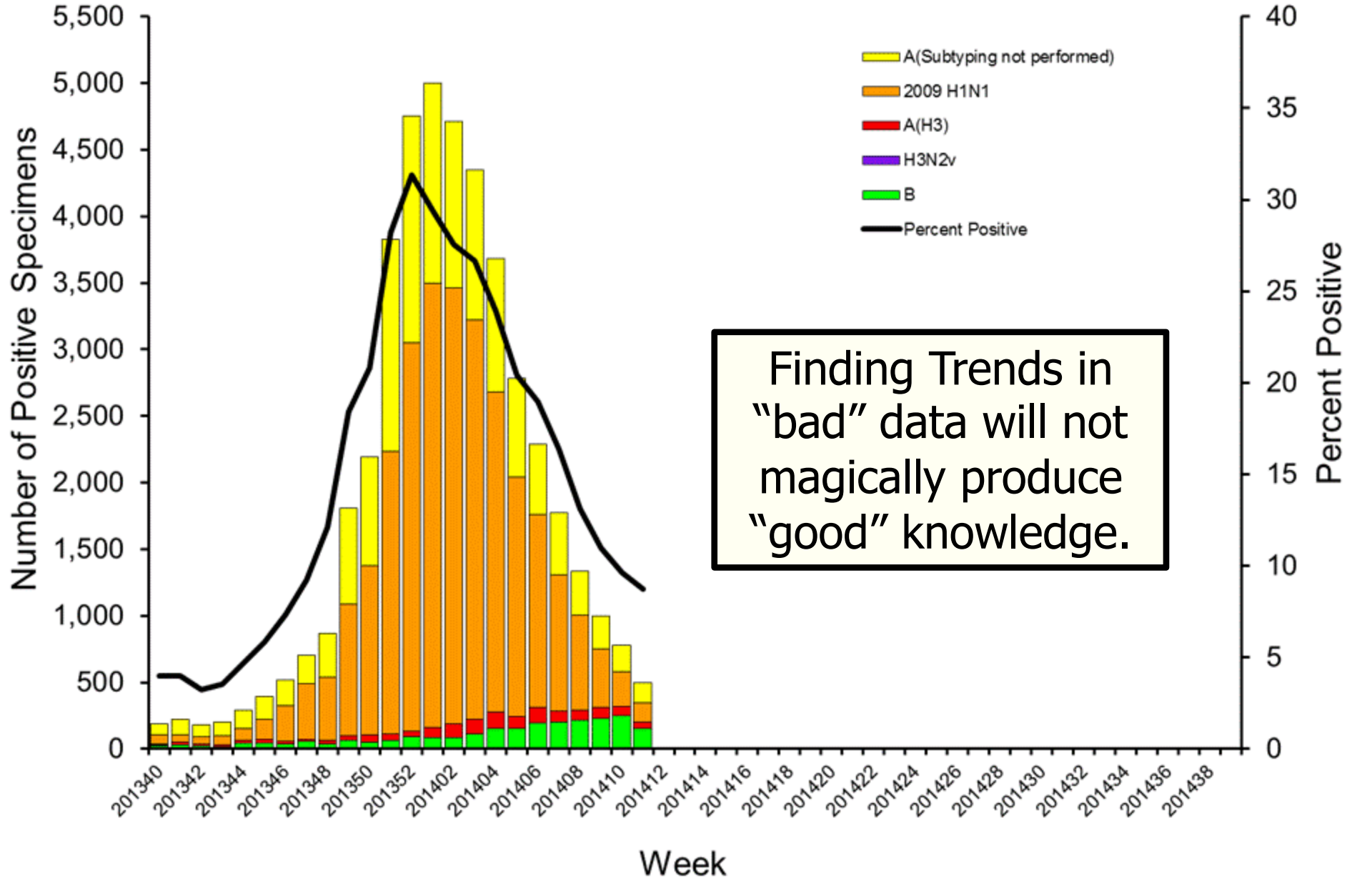
The overwhelming majority of people who think they have the flu, don't.

Influenza Positive Tests Reported to CDC by U.S. WHO/NREVSS Collaborating Laboratories, National Summary, 2013-14



So why would their search queries give you any indication of the incidence of flu?

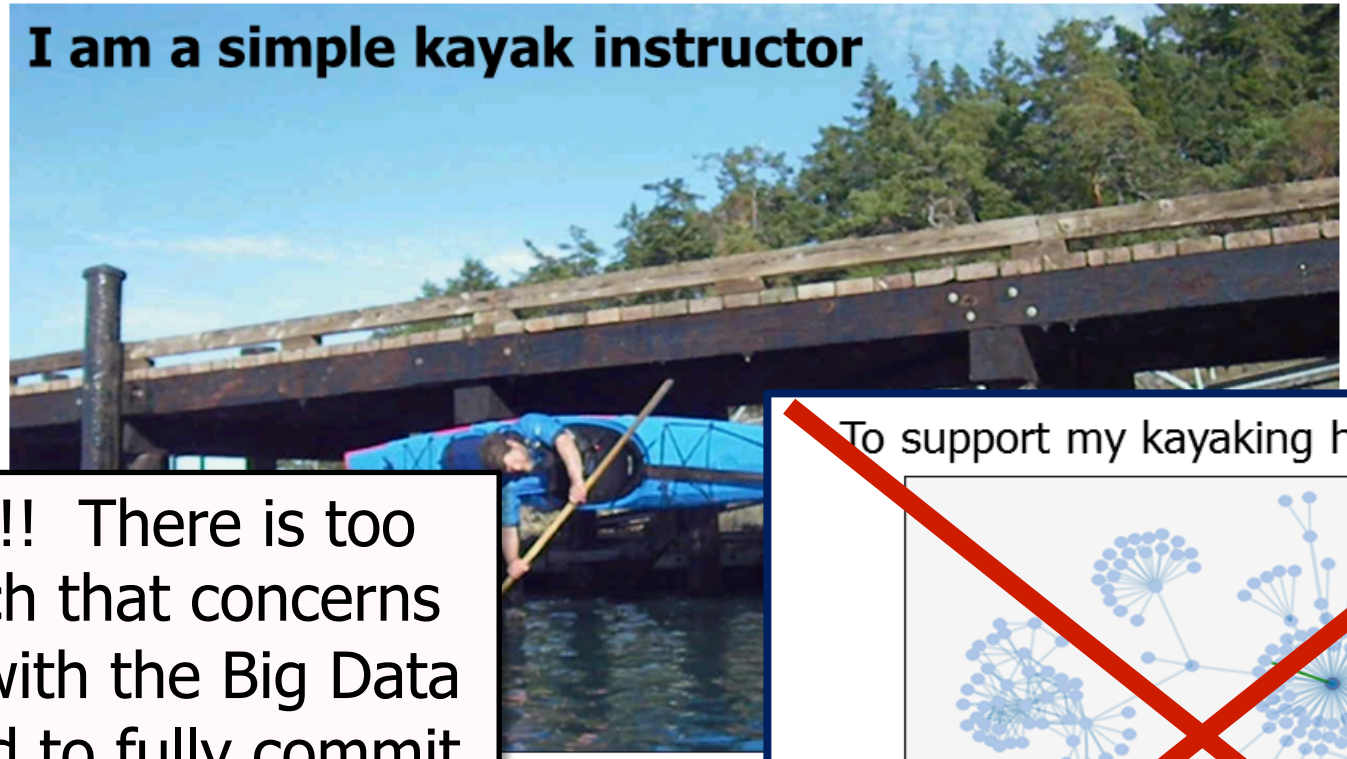
Influenza Positive Tests Reported to CDC by U.S. WHO/NREVSS Collaborating Laboratories, National Summary, 2013-14



Finding Trends in "bad" data will not magically produce "good" knowledge.

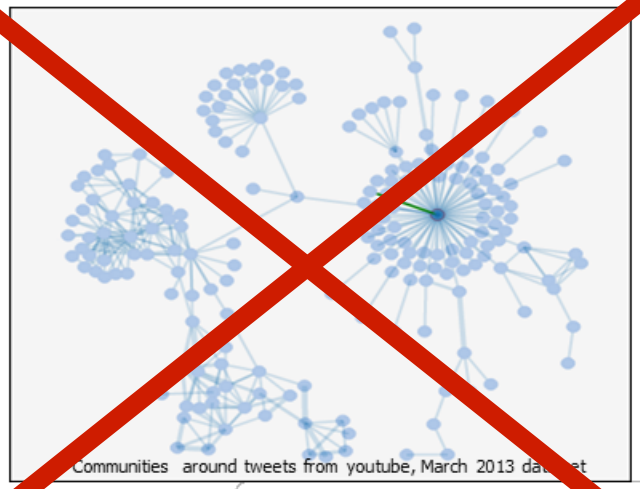
Am I ready to change my Introduction slide?

Introduction



NO!! There is too much that concerns me with the Big Data world to fully commit to the game.

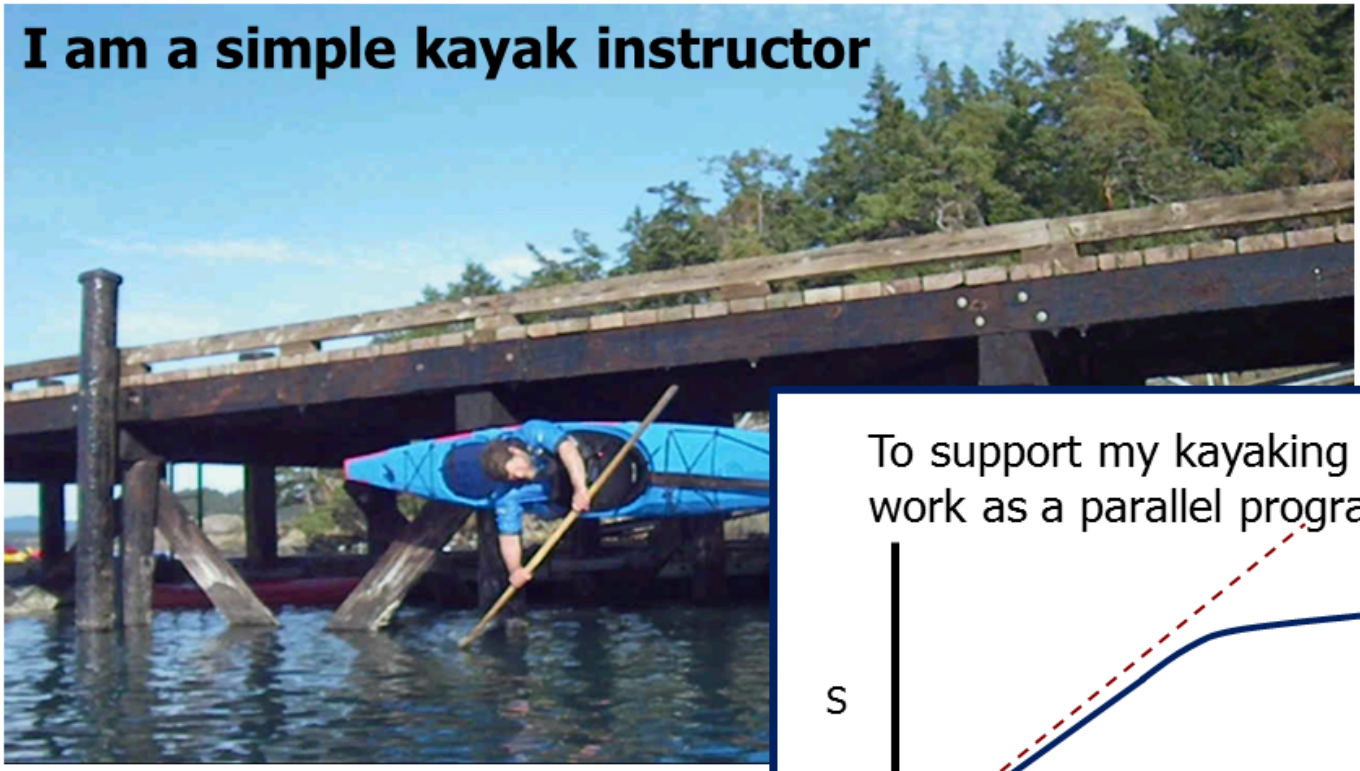
To support my kayaking habit



I create technologies that make people buy things they don't really need.

I think I'll stick to my old Introduction slide?

Introduction



To support my kayaking habit I work as a parallel programmer

Which means I know how to turn math into lines on a speedup plot

Summary

- Big Data Computing on real world workloads will require:
  - Computing where the data resides
  - Analytics integrated with the DBMS
  - Good usability so data scientists can get their job done
- At the Big Data Intel Science And Technology Center (BD-ISTC) we are working on a solution stack to address these and related topics.
- Stay tuned as we:
  - Address basic research questions on unified query languages and data-analytics/DBMS integration
  - Build a prototype to test our concepts (between now and 2017).